LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-653616

Polymorphic Peptide Hair Project

G. J. Parker, D. S. Anex, M. F. Leppert, L. Baird, N. Matsunami, T. Leppert

April 23, 2014

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Polymorphic Peptide Hair Project

Personnel:

Glendon Parker	Protein-Based Identification Technology
Deon Anex	Forensic Science Center, Lawrence Livermore National Laboratory
Mark Leppert	University of Utah
Lisa Baird	University of Utah
Nori Matsunami	University of Utah
Tami Leppert	Protein-Based Identification Technology

Project: 38667
Task: 6447253
Contract B601942

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Contents

Summary 1

Background 1

Objectives 3

Methodology 3

Results 4

- Proteomic Data Acquisition 4
- Peptides Containing Non-Synonymous SNPs Identified in the Hair Proteome 6
- Calculation of Power of Discrimination 6
- Calculation of Likelihood of Biogeographic Background 8
- Pair-Wise Linkage Disequilibrium Analysis 10
- Comparison of Proteomic Genotyping and Sanger Sequencing 11

Discussion 14

Future Work 14

References 15

Appendices 16

- Appendix 1. Biogeographic Background of Subjects using 129 racially informative SNPs 17
- Appendix 2. SOP: Hair Trypsinization Protocol 18
- Appendix 3. LC/MSMS Data Acquisition 20
- Appendix 4. Primer Sequences Used for Genotyping with Sanger Sequencing 23
- Appendix 5. Genotyping based on Sanger Sequencing 24
- Appendix 6. Pair-Wise Linkage Disequilibrium Analysis 25
- Appendix 7. Development of PBIT FASTA Database File 26
- Appendix 8. Peptide Selection Method 28
- Appendix 9. Overall Profile Frequency Recalculated with Strictest Product Rule Application 29
- Appendix 10. Primer / Factors for Consideration 30

Summary

This project achieves the empirical and conceptual foundation that allows for the forensic use of proteins to reveal genetic information and link informative samples with a given individual. The method is completely independent of DNA typing and DNA-based methodology. Information can now be obtained from hair, or potentially any protein source, when DNA-based methods are not feasible or provide incomplete information. This expands the amount of biological information that can be obtained in a forensic or intelligence context.

We have identified and characterized 60 peptides in human hair that contain genetic information in the form of non-synonymous single nucleotide polymorphisms (nsSNPs). In a study of 54 European American individuals, using 10 mg of hair, an average 745 ± 325 unique peptides were identified per run. An overall nsSNP profile frequency of up to 1 in 9000 was achieved in a single run with an average of 1 in 325 across the cohort. When the same calculations are conducted, but using the frequency of nsSNP peptide prevalence in the African population, the value changed to an average of 1 in 122,000. This change in calculated profile frequency (indicating a higher likelihood of European genetic origin rather than African) demonstrates the potential of a protein-based approach to distinguish genetic populations.

The ability of proteomic assignment to genotype nsSNP loci was directly validated by Sanger sequencing; 426 determinations were correctly made along with 7 misassignments (False Positive Rate (FPR) = 1.6%)). When one problematic peptide is excluded (rs11078993), the statistics improve to 4 misassignments (FPR = 0.93%). The total specificity of peptides identified as containing nsSNPs was 99.1%, with an overall sensitivity of 30.9%.

An important component for the use of genetics in biometrics is the use of the ”product rule”, or the creation of an overall probability that a given genetic profile is associated with an individual. For the rule to apply, the genotypic frequencies at different loci need to be independent of each other so they can then be multiplied together. This study has assumed that expressed genes can effectively be treated as single loci. To determine the appropriate boundaries beyond which independence can be assumed we conducted a pair-wise analysis of linkage disequilibrium. As expected by the clustering of some of the genes, some linkage between genes has been identified. This allows appropriate linkage boundaries to be drawn and correct employment of the product rule. If the most conservative reinterpretation of the product rule is applied (i.e., only one locus used per cluster) then the overall nsSNP profile frequency changes from a maximum of 1 in 9000 to a maximum of 1 in 785. On average, an individual profile frequency decreased by a factor of 4 when the most conservative calculation is applied. Due to lack of linkage disequilibrium between many common loci within the cluster, this interpretation is overly strict. Final values will be above this lower extreme value.

Background

There is a need to increase the number of biometric forensic methods that are quantifiable and increase the scope of samples that can be examined in a forensic and intelligence context[1]. Recent advances in DNA typing have greatly expanded the capacity to place individuals at a specific time and place, or associate them with a given artifact[2]. However, these methods depend on the presence of

sufficient DNA template to obtain identifying information and, unfortunately, DNA can be unstable and more easily degraded by environmental and biological processes[3, 4]. In the event that DNA typing yields an incomplete or null result, few quantifiable options are available to the investigator[1]. There is a need, therefore, to expand identifying technologies beyond those that depend purely on DNA typing[5].

This study investigates the feasibility of using genetic information retained in proteins to develop forensically relevant measures of identity. One particular type of genetic variation, non-synonymous single nucleotide polymorphisms (nsSNPs), is retained in proteins and thus can provide evidence of genetic variation after the original DNA template has degraded. The advantage of this approach is that protein is considerably more stable than DNA. The peptide backbone is very stable, while the phosphodiester bonds of DNA are more reactive to water and more susceptible to environmental changes, such as with pH or temperature[6]. Protein also occurs at much higher concentrations; put more simply, if you can observe something biological, what you see is mostly protein[5]. Proteomic methodologies, particularly tandem LC/MSMS, have developed to the point where minute amounts of material (< attomole) can provide a considerable amount of biological information[7, 8]. Large amounts of data can be generated using mass spectrometry. Recent proteomic analysis of a yeast extract, admittedly an ideal proteomic target, yielded over 46,000 unique peptides[9]. Structural disadvantages of using protein to obtain genetic information consist of three factors: a lack of a protein equivalent to PCR; nsSNPs are almost always bi-allelic and, therefore, less discriminate than the short-tandem repeat loci used in DNA typing; and, finally, each protein source or tissue examined only represents the subset of gene products that have been expressed in that tissue[10].

Several milestones need to be achieved to demonstrate that protein typing is feasible. There needs to be sufficient powers of discrimination so that relevant samples can confidently be assigned to a given individual, or at least provide a high confidence of exclusion. This means that sufficient numbers of genetic variation are identified and characterized, and the methodology needs to be sensitive enough to identify them in a forensic sample. The frequency of each locus, or groups of loci, needs to be independent so that the “Product Rule” can be employed. The method must be specific so that identified nsSNP-containing peptides accurately reflect the status of genetic variation in DNA. The method needs to be reproducible, robust, and widely applicable.

The origin of common nsSNPs (>5% of the population) pre-dates the exodus of humans out of Africa[11]. However due to genetic factors such as selection, founder effects, bottles necks, genetic drift, and admixture; the allelic frequency of each nsSNP can vary dramatically as a function of biogeographic background. Because of this it is theoretically possible to make inferences about the genetic background of an individual based on the allelic frequencies with which a given nsSNP occurs in a given population. That is, if an nsSNP identified in an individual is common in one population and rare in a second, it is more likely that the individual is genetically related to the first population.

To test the feasibility of protein typing we have chosen to focus on hair. Hair is a frequent, if under-utilized, component of crime scenes; an individual sheds about 100 to 150 per day[12]. Current forensic use of hair samples for identification is generally limited to mitochondrial DNA or visual

comparisons based on characteristics such as color, size and morphology. The protein in the hair represents a rich and untapped source of forensic information. A single hair is genetically discrete, so no mixed contributions occur. A hair shaft is highly robust, persisting in the environment beyond the loss of other tissues to decay, with the obvious exception of bone and teeth[13, 14]. As a practical matter it is also easy to collect. Several studies have shown that the proteomic profile of hair is quite complex, containing house-keeping proteins, and cellular components in addition to the more documented (and abundant) trichocyte keratins and their associated proteins[15, 16]. This means that the proteome of hair provides a sufficient number of proteins to make genetic calculations of identity and background.

Objectives

To test the feasibility of using protein as a source of identifying genetic information we pursued the following goals:

A) Identify and characterize nsSNP-containing peptides in the proteomic datasets of hair digests.
B) Calculate the power of discrimination that can be obtained from conducting a proteomic analysis of hair from an individual.
C) Compare the sample to a European or African background, based on differing frequencies with which a peptide occurs in each population.
D) Verify proteomic analysis using DNA sequencing analysis for each individual.
E) Determine the specificity and sensitivity of identified nsSNP-containing peptides.
F) Conduct a pair-wise linkage disequilibrium analysis of identified nsSNPs to determine the extent to which the "product rule" can be employed.

Methodology

To meet the above objectives we obtained complex proteomic datasets from hair, identified nsSNP-containing peptides in the peptide mixtures, and then conducted genetic analyses on identified peptides. We calculated both nsSNP-profile frequencies and likelihood of biogeographic background. To calculate sensitivity and specificity we then confirmed the assignment of mass spectra corresponding to nsSNPs by directly sequencing the loci within the DNA of each subject. We also analyzed each nsSNP locus, when present in the haplotype databases, for pair-wise linkage disequilibrium to determine the appropriate application of the product rule.

Hair and DNA were collected from 60 individuals, who self-reported for European genetic background (L1.001 to L1.060; Sorenson Forensics, LLC; Salt Lake City, UT). Hair was milled, extracted, reduced and carboxymethylated, and digested with trypsin (see, Appendix 2). The resulting complex peptide mixture was analyzed by liquid chromatography/mass spectrometry (LC/MS) using an Agilent 6530 Accurate-Mass Q-TOF LC/MS mass spectrometer (Appendix 3). The resulting datasets were converted to "mgf" format using Mass Hunter Workstation Software (Version B.06.00, Build 6.0.633.0, Agilent Inc.) (Appendix 3).

The converted "mgf" files were analyzed by two methods. The first method used the GPM manager (www.thegpm.org) that utilized a crowd-sourced reference database (GPMdB). The second method used X!Tandem (Trans Proteomic Pipeline) with a custom variant database (developed by Tami Leppert, Salt Lake City, UT, Appendix 7) designed to include all possible nsSNPs with an allelic frequency

above 0.5%, in either the European or African population. The redundant analysis balanced accessibility and completeness. The GPM-manager was more accessible, but the variant database was incomplete resulting in less nsSNP identifications. The custom database was complete, but also less user friendly. The different approaches were expected to yield overlapping datasets.

Peptides containing nsSNPs that were identified in the GPM manager analysis were screened for proteomic quality, uniqueness, and allelic frequency (>1%). When necessary, mass spectra were manually examined to determine if there were fragmentation masses consistent with the alternate reference allele, in which case the peptide was most likely a misassignment, or whether any masses were due to other species (**Table 1**). Once identified, the peptides were collated for each individual and the genotype at each nsSNP locus imputed (**Table 3**). Peptides from different loci were multiplied together using the frequencies with which the peptide occurs in the European population to provide a combined nsSNP profile frequency (**Figure 1**). These calculations assumed Hardy-Weinberg Equilibrium outside of the gene boundary and full linkage disequilibrium within it; multiple loci within a gene boundary were effectively treated as a single locus. The calculations were repeated using frequencies with which the peptides occurred in the African population (**Figure 3A**). The quotient of the overall profile frequency using European over African values provides a likelihood measure that the protein sample comes from a European background as opposed to the African population (**Figure 3B**).

The identification of nsSNP-containing peptides was directly validated by sequencing the loci in subject DNA using Sanger sequencing (Appendix 4, 5). Type-I (misassignment) and Type-II errors (missed assignment) were identified (**Table 3**) and the sensitivity and specificity of each nsSNP-containing peptide calculated (**Table 4**). Pair-wise linkage disequilibrium calculations were also conducted on nsSNP loci when the nsSNPs occurred in the haplotype databases (**Table 2**).

Results

Proteomic Data Acquisition

Hair from subjects (10.5 ± 0.5 mg) was processed physically and biochemically as described below (Appendix 2). A total of 54 individuals were studied: 3 from a previous cohort pre-dating this contract (U3, U5, and U17) and 51 from the subject cohort recruited as part of this study (L1.001 to L1.060). Nine subjects were excluded due to some admixture from other genetic groups identified through a test of genetic background based on 129 SNPs (Sorenson Genomics, LLC; Salt Lake City, UT). A standard and complex peptide mixture from individual U3 was applied 14 times across the series of sample sequences to determine consistency. There was a wide variation in the peptide yield across the cohort. When 20 µL of sample was injected on the LC/MS, the number of unique identified peptides ranged from 190 to 1244 (average = 724 ± 305 (stdev), median = 692). The average false discovery rate, generated by the GPM manager, of peptide spectra assignment was 1.1 ± 0.1% (average ± standard deviation) when using the GPM manager.

Table 1. Peptides Containing Non-Synonymous SNPs Identified in the Hair Proteome. Peptides bearing nsSNPs (peptide) in the hair proteome are listed in order of Gene Name (GN) below. The frequency of peptide prevalence for each allele in the European (EUR) and African (AFR) is indicated. The nsSNP is indicated in the peptide sequence in red, with the minor allele indicated in lower case (peptide). Genotypes later confirmed directly by Sanger sequencing are indicated (Seq), along with the chromosome (Chr) and whether the peptide was detected in this project (Proj).

	Frequency							
GN	EUR	AFR	nuc	peptide	rs#	Seq	Chr	Proj
BLMH	0.562	0.378	T	HVPEEVLAVLEQEPIvLPAWDPMGALA-	rs1050565	y	17	y
	0.884	0.927	A	HVPEEVLAVLEQEPIILPAWDPMGALA-	rs1050565	y	17	y
COL4A4	0.821	0.581	A	GPSGPGCKGEPGLDGR	rs3752895		2	y
	0.691	0.87	G	GPpGPGCKGEPGLDGR	rs3752895		2	y
DSP	0.269	0.642	A	GqSEADSDKNATILELR?rr5EADSDKNATILELR	rs6929069	y	6	
	0.984	0.858	G	GRSEADSDKNATILELR?	rs6929069	y	6	
GSDMA	0.812	0.915	T	QLNPqGDLTPLDSLIDFK?	rs3894194	y	17	
	0.491	0.642	C	GDLTPLDSLIDFK	rs3894194	y	17	
GSDMA	0.784	0.894	C	GHEVTLEALPK	rs56030650	y	17	
	0.699	0.476	A	GHEVnLEALPK	rs56030650	y	17	
GSDMA	0.731	0.227	G	ALETVQER	rs7212938	y	17	y
	0.781	0.979	T	ALETIQER	rs7212938	y	17	y
GSTP1	0.544	0.712	G	YvSLIYTNYEAGKDDYVK?	rs1695	y	11	
	0.905	0.826	A	YISLIYTNYEAGKDDYVK?	rs1695	y	11	y
HEXB	0.971	1	A	GILIDTSR?	rs10805890	y	5	y
	0.264	0.028	G	GILVDTSR?	rs10805890	y	5	
HIST1H1T	0.712	0.899	G	MSETVPAASASAGVAAMEKLPTK	rs198844		6	
	0.792	0.582	C	MSETVPAASASAGIAAMEKLPTK	rs198844		6	
JUP	0.121	0.02	T	SAIVHLINYQDDAELAThALPELTK?	rs41283425?	y	17	y
	1	1	C	SAIVHLINYQDDAELATR	rs41283425?	y	17	y
JUP	1	1	C	NEGTATYAAAVLFR??	rs143043662	y	17	y
	0.013	0	T	NEGTATYAAAiLFR??	rs143043662	y	17	
KRT1	0.042	0	A	NKLNDLEDALQQsKEDLAR	rs17678945	y	12	
KRT31	0.169	0.065	A	DNvELENLIR	rs6503627?	y	17	y
KRT31	0.05	0.004	A	SQYEvLVETNR?	rs112544857	y	17	y
KRT32	0.748	0.493	T	ADLEAQVEyLK	rs2071561	y	17	
KRT32	0.832	0.796	G	LEGEINTYR?	rs2071563	y	17	y
	0.682	0.663	A	LEGEINmYR?	rs2071563	y	17	y
KRT32	0.997	1	C	YSSQLAQMQCMITNVEAQLAEIRADLERQNQEYQVLLDVR	rs11078993	y	17	y
	0.09	0.085	T	YSSQLAQMQCMITNVEAQLAEIqADLERQNQEYQVLLDVR	rs11078993	y	17	y
KRT32	0.939	1	C	CQYEAMVEANRR	rs72830046?	y	17	y
	0.459	0.098	T	CQYEAMVEANhR	rs72830046?	y	17	y
KRT33A	0.446	0.597	A	QVVSSSEQLQSYQvEIIELR?	rs12937519	y	17	y
KRT34	0.575	0.089	A	SQYEALVEINR	rs2239710	y	17	y
KRT35	0.778	0.927	G	VSAMYSSSpCKLPSLSPVAR	rs743686	y	17	y
	0.702	0.451	A	VSAMYSSSSCKLPSLSPVAR	rs743686	y	17	y
KRT35	0.708	0.623	C	TNCSaRPICVPCPGGRF?	rs2071601	y	17	y
	0.798	0.859	G	TNCSPRPICVPCPGGRF?	rs2071601	y	17	y
KRT35	0.958	0.988	C	TNCSPRPICVPCPGGRF?	rs12451652	y	17	y
	0.306	0.134	T	TNySPRPICVPCPGGRF?	rs12451652	y	17	
KRT37	0.934	0.963	C	TSFYSTSSCPLGCTMAPGAR	rs9910204	y	17	y
	0.475	0.297	A	TSFYSTSSCPLcCTMAPGAR	rs9910204	y	17	y
KRT40	0.433	0.11	G	TASALEIELQAQQSLTESLECTVAETEAQYSSQLAQIQr?/?LIDNLENQLAEIR	rs150812789	y	17	y
KRT81	0.282	0.045	A	LYEEEILILQSHISDTSVVVK??	rs6580873	y	12	y
	0.56	0.122	C	GLTGGFGSHSVCr	rs2071588	y	12	y
KRT83	0.863	0.768	C	DLNMDCIVAEIK	rs2852464	y	12	y
	0.62	0.728	G	DLNMDCmVAEIK	rs2852464	y	12	y
KRT84	0.942	0.996	G	rEYQELMNAKLGLDIEIATYR?	rs951773	y	12	
	0.48	0.167	A	CEYQELMNAKLGLDIEIATYR??	rs951773	y	12	
KRT85	1	1	G	IAVGGFRAGSCGR?/?AGSCGR	rs61630004	y	12	y
	0.092	0.041	A	IAVGGFRAGSCGhSFGYR?/?AGSCGhSFGY	rs61630004	y	12	y
KRTAP10-8	0.095	0.089	A	TYVIAASTMSVCSSDVGHVSR	rs411254		21	
	0.997	1	G	TYVIAASTMSVCSSDVGr	rs411254		21	
LRRC15	0.395	0.154	A	ELSIGIFGPMPNLR	rs13070515?	y	3	y
	0.944	0.984	G	ELSPGIFGPMPNLR	rs13070515?	y	3	y
LRRC15	0.937	0.955	C	LYLSNNHISQLPPSVFMQLPQLNR	rs13060627	y	3	
	0.443	0.374	T	LYLSNNHISQLPPSiFMQLPQLNR	rs13060627	y	3	
NEU2	0.755	0.973	C	EWSTFAVGPGHCLQLHDR	rs2233391		2	y
	0.778	0.285	A	EWSTFAVGPGHCLQLnDR	rs2233391		2	y
S100A3	0.998	0.858	C	ARPLEQAVAAIVCTFQEYAGR	rs36022742	y	1	y
	0.085	0.589	T	AkPLEQAVAAIVCTFQEYAGR	rs36022742	y	1	y
SERPINB5	0.52	0.74	C	GVALSNVvHK?	rs1455555	y	18	
	0.731	0.297	T	GVALSNVIHK?	rs1455555	y	18	
TGM3	0.982	0.838	T	AALGVQSINWQkAFNR	rs214803	y	20	y
	0.317	0.696	G	AALGVQSINWQTAFNR	rs214803	y	20	y

Data sets were converted to the mass spectrometry generic format (mgf) and analyzed using two different approaches: by submission to the GPM manager software (www.thegpm.org, release SLEDGEHAMMER (2013.09.01)) and to the Petunia Graphic user interface (TANDEM CYCLONE TPP, download = 2011.12.01.1 - LabKey, Insilicos, ISB, Appendix 7)). The GPM manager uses a crowd-sourced GPM database and the Petunia GUI used a custom protein reference database that was developed as part of the project (Tami Leppert, Salt Lake City, Appendix 8).

Peptides Containing Non-Synonymous SNPs Identified in the Hair Proteome

Peptides containing nsSNPs (rs#) in proteomic datasets from hair preparations were identified using Peptide Spectral Matching software and collated according to Gene Name (GN). For each allele, both nucleotide (nuc) and peptide sequence (peptide) is represented, with the amino acid change represented in red and the minor allele in lower case (**Table 1**). Peptides that could be assigned to two or more genomic locations were excluded. The percentage of the European (EUR) and African (AFR) population containing each allele was calculated using data from the 1000 Genomes project (combined homozygote and heterozygote frequencies). Peptides were filtered by the following criteria: each peptide sequence could only occur at one locus in the genome, each spectrum had to have a suitable expectation score (log(e) < -2, X!Tandem), and the minor allelic frequency had to be greater than 1% in either of the reference genetic populations. Resulting spectra were manually examined to determine if any masses were consistent with the alternate allele, in which case the most likely scenario was a mass shift due to chemical modification of the peptide.

Chromosomal location is indicated (Chr) to illustrate loci that uniquely reside on different chromosomes and, therefore, are independent. The nsSNP-containing peptides were directly confirmed by Sanger sequencing of the subject's DNA (Seq) (**Table 3**). Peptides that derive from data predating the contract are included, with those detected in this dataset indicated (Proj) in **Table 1**.

Calculation of Power of Discrimination

Peptides containing nsSNPs were identified, and collated for each subject. The frequencies of each peptide in the European population of the 1000 Genomes project were calculated and multiplied together to create a combined profile frequency (1 in; $1/\Pi f_{EUR}$; **Figure 1**). Hardy-Weinberg equilibrium was assumed for loci outside of the gene boundary and complete disequilibrium was assumed for loci within a gene boundary. Haplotype frequencies were estimated for loci located within gene boundaries (see below). The precise boundaries defining areas of linkage disequilibrium can now begin to be more precisely drawn as a result of the pair-wise linkage disequilibrium analysis conducted below (**Table 2** and Appendix 6).

Power of genetic discrimination scores ranged from 1 in 1.002 (subject L1.016) to 1 in 9000 (subject L1.058). Hair from individual U3 was used as a standard. The average Power of Discrimination was 1 in 280.

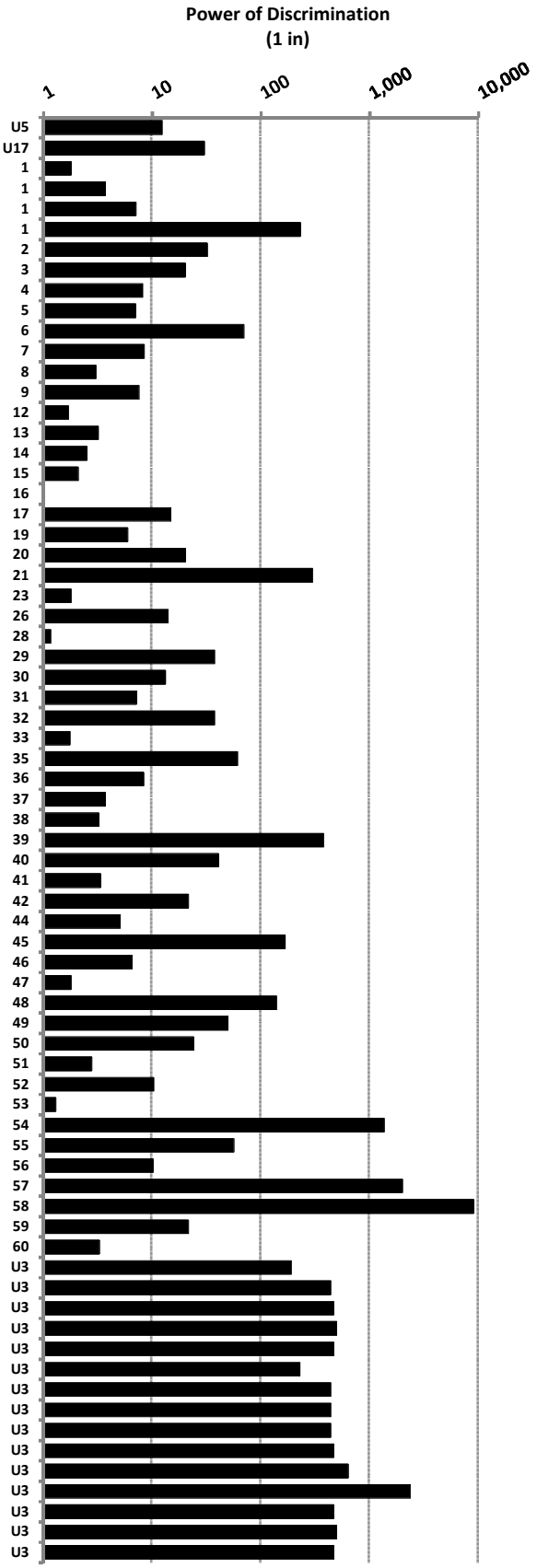


Figure 1. Overall nsSNP Profile Frequency

The Power of Discrimination improves as a function of proteomic quality, measured as the number of unique peptides identified in the dataset (**Figure 2**). After conversion to logarithmic values, linear regression (dashed line; y = 0.0024x-0.3023) resulted in a significant correlation with an r^2 value of 0.624 (n = 58, p < 0.0001). Values corresponding to 20 µL of the standard (U3) are indicated (red).

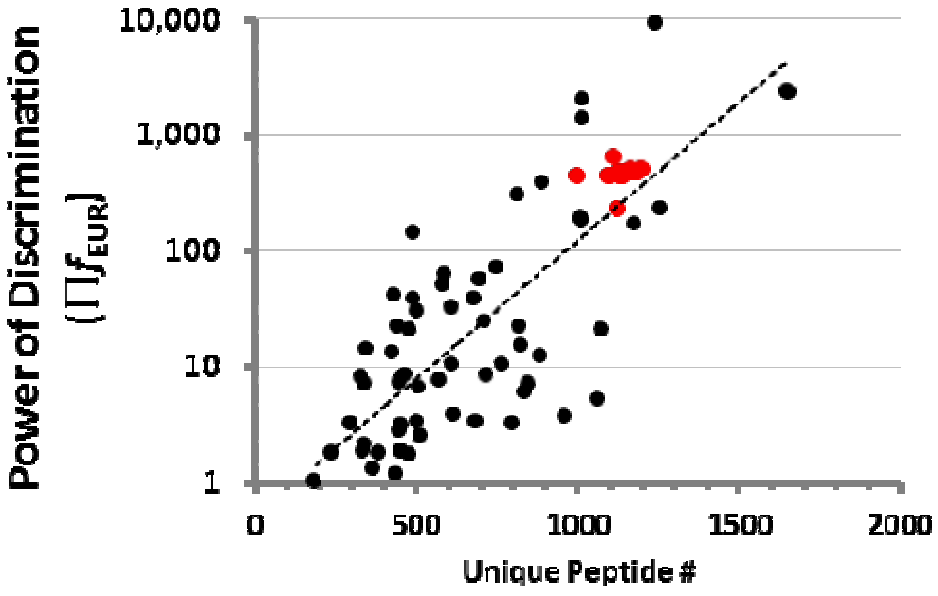


Figure 2. Power of Discrimination as a Function of Proteomic Dataset Quality.
Genetic power of discrimination was calculated for each individual in the cohort and plotted against the corresponding number of unique peptides in the proteomic dataset (black circles). Data corresponding to 20 µL of the standard hair peptide mixture are indicated (red circles).

Calculation of Likelihood of Biogeographic Background

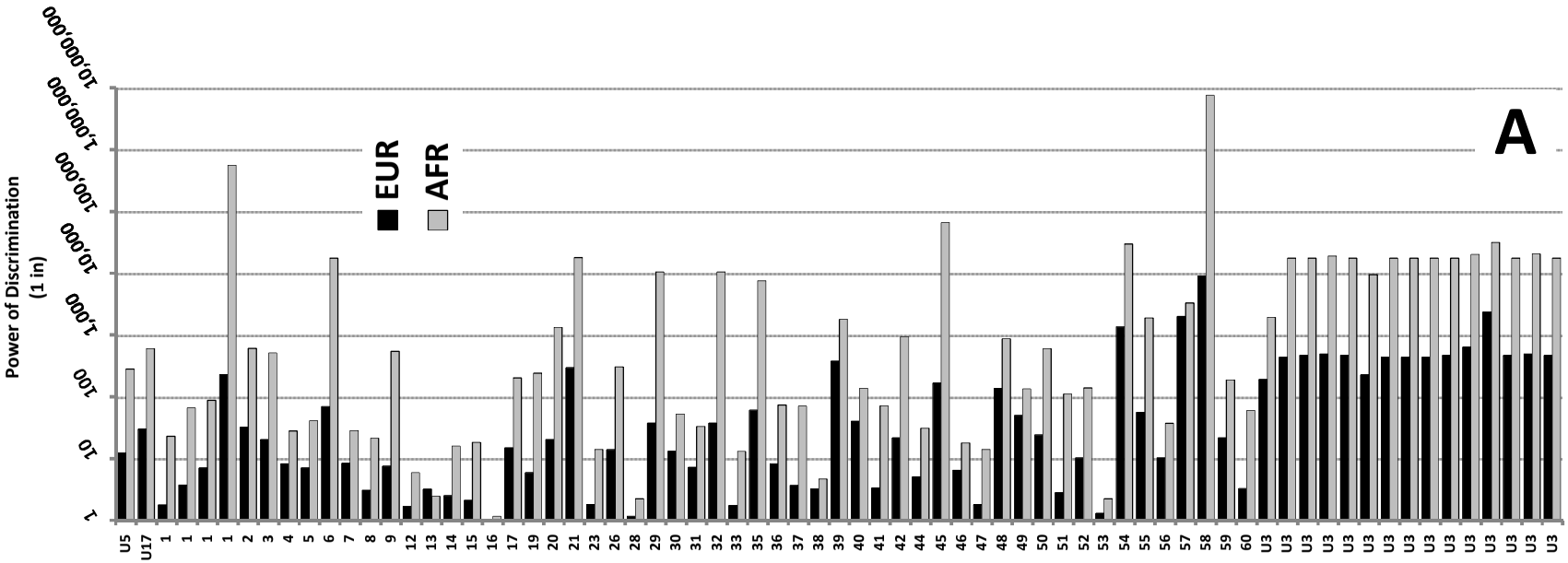
Individual nsSNP frequencies vary among different genetic populations. As such, nsSNPs may be used to identify an individual's likely biogeographic background. When the Power of Discrimination is calculated using peptide frequencies in the African, as opposed to the European, population, higher values are obtained (**Figure 3A**). The quotient of these values can be used as a likelihood measure that compares the hair sample to one population relative to another, in this case European relative to African biogeographic background (**Figure 3B**).

Across the cohort the likelihood measures range from 1 to 780 with an average of 50, a median of 18, and a standard deviation of 116. (n = 64) These values show a much higher likelihood of the samples originating in the European genetic population relative to the African genetic population. This is consistent with the study design (subjects selected to have a European genetic background). The average 50-fold differential obtained in this study should be taken as an indicator. This is not a probability measurement, but rather a comparison of each sample to one population and another. Increased identification of nsSNP loci will naturally increase this differential and increase the confidence in assigning a biogeographic background to a particular sample.

Figure 3. Comparison of Hair Samples to the European and African Population

A) The overall profile frequency of identified nsSNPs per individual was calculated using the frequencies (combined heterozygote and homozygote) with which the nsSNPs occur in the European (EUR) population (black bars, **Figure 1**.). The same calculations were conducted using the frequency with which the nsSNPs occurred in the African (AFR) population (grey bars).

B) The quotient of the overall nsSNP profile frequency (Π*f*) using European and African frequencies provides a likelihood measure associating individual hair samples with the European and African genetic population.



Pair-Wise Linkage Disequilibrium Analysis

Appropriate application of the "Product Rule" requires that the boundaries of regions of linkage disequilibrium are determined empirically through a pair-wise analysis for each nsSNP identified in this study. Many nsSNP loci occur throughout the genome on different chromosomes (**Table 1**). However the more abundant keratin genes tend to cluster in two chromosomes: acidic Type-I keratins (KRT31 to KRT38) on chromosome 17 and basic Type-II keratins (KRT81 to KRT85) on chromosome 12[17, 18]. This raises the possibility of pair-wise linkage disequilibrium between other genes in the cluster, which would prevent a blanket use of the product rule. To test this we conducted a pair-wise linkage disequilibria analysis on nsSNPs identified in the proteomic datasets using the SNP Annotation and Proxy Search tool from the Broad Institute (http://www.broadinstitute.org/mpg/snap/ldsearchpw.php) (**Table 2**). D' values were calculated with complete linkage (D' = 1) formatted in red, and no linkage (D' = 0) formatted in green. As seen in chromosome 17 there is very little linkage between KRT32 and KRT34 (D' = 0.06), and KRT32 and KRT35 (D' = 0.163). However, there is a high level of linkage between KRT34 and KRT35 (D' = 0.863), indicating that the product rule should not be used if these two loci are in a dataset. Instead they should perhaps be treated as a single haplotype. The nsSNPs observed in the datasets with KRT81 (detected once in individual L1.058) and KRT82 have a pair-wise disequilibria value of D' = 0.689, which is an intermediate value. A more complete analysis, including nsSNP loci that are discovered from previous analyses and not identified and confirmed in this dataset, is enclosed below (Appendix 6).

CHR 17 D' values

		MAF (EUR)	0.344	0.426	0.462
			KRT34	***KRT32***	***KRT35***
MAF (EUR)		**rs#**	**rs2239710**	**rs2071563**	**rs743686**
0.344	***KRT34***	**rs2239710**		0.06	0.803
0.426	***KRT32***	**rs2071563**	0.06		0.163
0.462	***KRT35***	**rs743686**	0.803	0.163	

CHR 12 D' values

		MAF (EUR)	0.152	0.379	0.269
			KRT81	***KRT83***	KRT84
MAF (EUR)			**rs6580873**	**rs2852464**	**rs951773**
0.152	***KRT81***	**rs6580873**		0.689	0.058
0.379	***KRT83***	**rs2852464**	0.689		0.863
0.269	KRT84	**rs951773**	0.058	0.863	

Table 2. Pair-Wise Linkage Calculations of nsSNP Loci.

Blocks of linkage occur in the genome as a function of allelic frequency and distance between loci. Both a short distance and a shorter amount of genetic time, which is the usual case for rarer alleles, reduce the likelihood that recombination would separate loci, eventually result in Hardy-Weinberg equilibrium, and allow use of the product rule. The clustering of keratin genes on chromosome 12 and 17, and evident linkage disequilibrium for some loci within the cluster, means that some evaluation of appropriate boundaries of linkage disequilibrium needs to take place. A strict, and overly conservative interpretation, which ignores examples of pair-wise independence, is included in Appendix 10. We are consulting on this issue with Dr. Bruce Weir of the University of Washington, who is perhaps the foremost expert forensic genetics in the country.

Comparison of Proteomic Genotyping and Sanger Sequencing

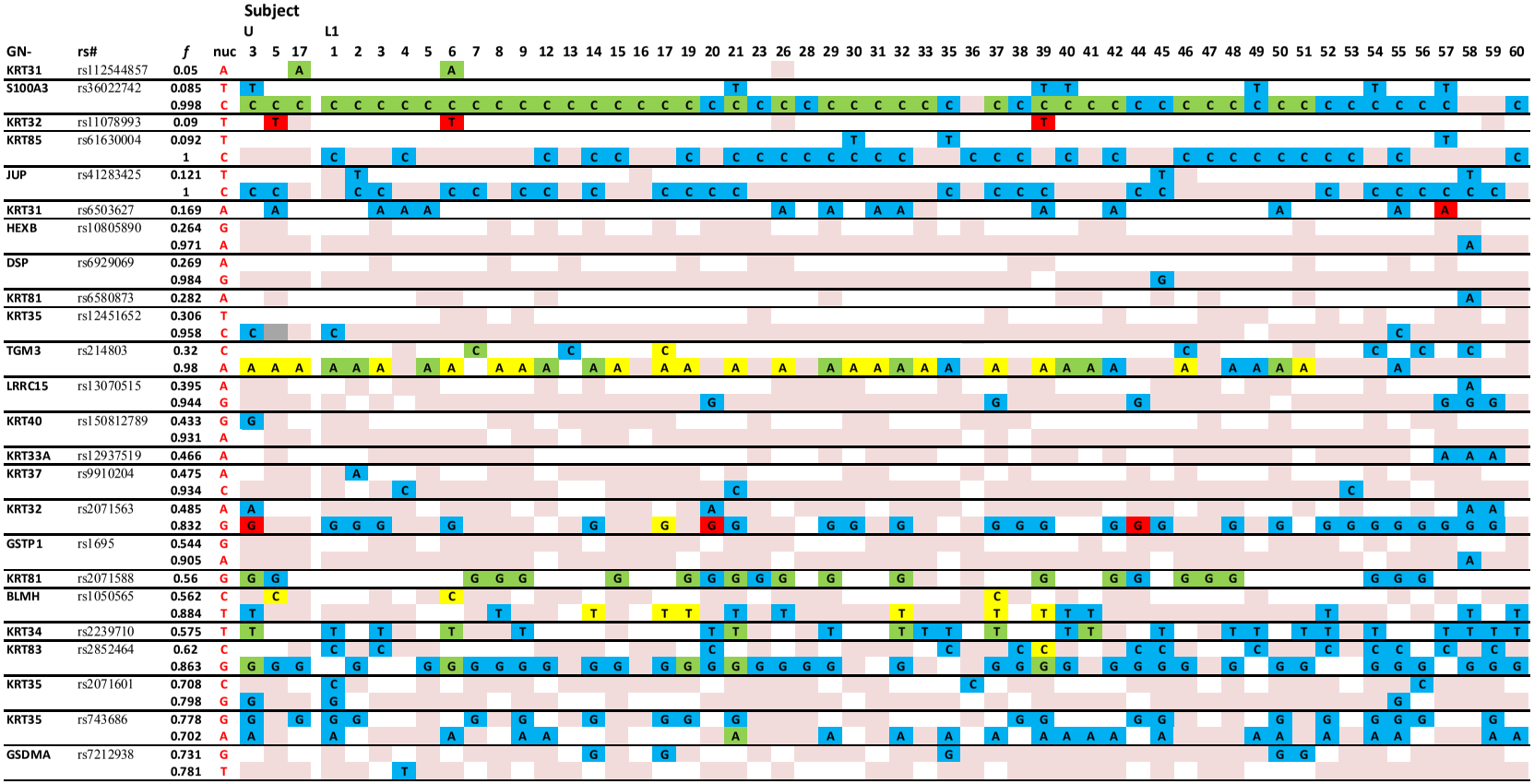
For protein-based genotyping to be feasible it needs to be specific, with peptide sequence assignment reflecting the presence of variants within the genome. Peptides containing nsSNPs (rs#) were identified in proteomics datasets using redundant mechanisms: GPM manager software (www.thegpm.org) and the Trans Proteomic Pipeline (tools.proteomecenter.org). The imputed genotypes based on these two approaches were then compared with the direct approach based on Sanger Sequencing of the subjects DNA (**Table 3**). Identified nsSNP-containing peptides from genes (GN) were collated for each individual and genotypes were assigned based on the nucleotide responsible for each alternative peptide (nuc). Assigned genotypes were compared to direct Sanger sequencing of the loci in the DNA of each individual (Appendix 4, 5). Failed amplification is indicated by grey. The frequency of each variant within the European population is indicated (*f*).

A total of 433 genotype determinations were made from the proteomic data. When compared with Sanger sequencing 426 assignments were confirmed (overall specificity = 98.4%) and 7 assignments were shown to be incorrect (red squares; False Positive Rate (FPR) = 1.61%). When a problematic peptide (rs11078993; specificity = 0%, **Table 4.**) was removed from the dataset the FPR was reduced to 0.93% (specificity = 99.1%). The overall sensitivity (ratio of the number of found peptides to the number of expected peptides based on the Sanger sequencing) of identified peptides containing nsSNPs was 30.9%. Higher levels of sensitivity will occur when a greater proportion of peptides a complex mixture can be identified[7, 9]. A detailed compilation of the performance of individual peptides is provided in **Table 4**. The four instances of Type-I error will require a repeated sample preparation and data acquisition, along with repeated Sanger sequencing, to determine if they are repeatable.

Table 3. Direct Comparison of Proteomic Genotyping with Sanger Sequencing of Subject DNA.

Identified nsSNP-containing peptides (rs#) from genes (GN) were collated for each individual and genotypes assigned based on the nucleotide responsible for each alternative peptide (nuc). Assigned genotypes were compared to direct Sanger sequencing of the loci in the DNA of each individual (Appendix 4, 5). Datasets from 2 populations were examined, a selection of three from a previous dataset (U3, U5, U17) and 51 from a cohort of European Americans (L1.001 to L1.060). Peptides identified using the GPM manager are indicated with blue, those identified using the TPP indicated in yellow and redundant peptides identified in green. Type-I errors are indicated by red.

To increase confidence that the proteomic assignment is correct, acceptance criteria were employed to screen Peptide Spectral Matches: proteomic measures of quality were employed (log(e) < -2, GPM manager; Expect score < 1, Trans Proteomic Pipeline), only peptides associated with common alleles were used to maximize the probability that mass shifts occurring in given peptides are due to genetic mechanisms and not chemical modification (allelic frequency > 1%). Finally, spectra corresponding to nsSNP-containing peptides were manually inspected to ensure that fragmentation masses consistent with the alternative allele or other


Subject
U
L1
GN-
rs#
f
nuc
KRT31 rs112544857 0.05 A
S100A3 rs36022742 0.085 T
0.998 C
KRT32 rs11078993 0.09 T
KRT85 rs61630004 0.092 T
1 C
JUP rs41283425 0.121 T
1 C
KRT31 rs6503627 0.169 A
HEXB rs10805890 0.264 G
0.971 A
DSP rs6929069 0.269 A
0.984 G
KRT81 rs6580873 0.282 A
KRT35 rs12451652 0.306 T
0.958 C
TGM3 rs214803 0.32 C
0.98 A
LRRC15 rs13070515 0.395 A
0.944 G
KRT40 rs150812789 0.433 G
0.931 A
KRT33A rs12937519 0.466 A
KRT37 rs9910204 0.475 A
0.934 C
KRT32 rs2071563 0.485 A
0.832 G
GSTP1 rs1695 0.544 G
0.905 A
KRT81 rs2071588 0.56 G
BLMH rs1050565 0.562 C
0.884 T
KRT34 rs2239710 0.575 T
KRT83 rs2852464 0.62 C
0.863 G
KRT35 rs2071601 0.708 C
0.798 G
KRT35 rs743686 0.778 G
0.702 A
GSDMA rs7212938 0.731 G
0.781 T

GN	rs#	*f*	nuc	peptide	sensitivity		specificity
KRT31	rs112544857	**0.05**	**A**	SQYEvLVETNR	**67%**	(2/3)	**100%**
S100A3	rs36022742	**0.085**	**T**	AkPLEQAVAAIVCTFQEYAGR	**100.0%**	(7/7)	**100%**
S100A3	rs36022742	**0.998**	**C**	ARPLEQAVAAIVCTFQEYAGR	**94.4%**	(51/54)	**100%**
KRT32	rs11078993	**0.09**	**T**	YSSQLAQMQCMITNVEAQLAEIhADLERQNQEYQVLLDVR	**0.0%**	(0/3)	**0%**
KRT85	rs61630004	**0.092**	**T**	IAVGGFRAGSCGR / AGSCGR	**100.0%**	(3/3)	**100%**
KRT85	rs61630004	**1**	**C**	IAVGGFRAGSCGhSFGYR / AGSCGhSFGY	**53.7%**	(29/54)	**100%**
JUP	rs41283425	**0.121**	**T**	SAIVHLINYQDDAELAThALPELTK	**50.0%**	(3/6)	**100%**
JUP	rs41283425	**1**	**C**	SAIVHLINYQDDAELATR	**48.1%**	(26/54)	**100%**
KRT31	rs6503627	**0.169**	**A**	DNvELENLIR or QLERDNvELENLIR	**93.3%**	(14/15)	**93%**
HEXB	rs10805890	**0.264**	**G**	GILVDTSR	**0.0%**	(0/18)	**na**
HEXB	rs10805890	**0.971**	**A**	GILIDTSR	**1.9%**	(1/54)	**100%**
DSP	rs6929069	**0.269**	**A**	GqSEADSDKNATILELR / SEADSDKNATILELR	**0.0%**	(0/11)	**na**
DSP	rs6929069	**0.984**	**G**	GRSEADSDKNATILELR	**1.9%**	(1/53)	**100%**
KRT81	rs6580873	**0.282**	**A**	LYEEEILILQSHISDTSVVVK	**7.7%**	(1/13)	**100%**
KRT35	rs12451652	**0.306**	**T**	TNySPRPICVPCPGGRF-	**0.0%**	(0/14)	**na**
KRT35	rs12451652	**0.958**	**C**	TNCSPRPICVPCPGGRF-	**5.8%**	(3/52)	**100%**
TGM3	rs214803	**0.32**	**C**	AALGVQSINWQTAFNR	**63.6%**	(7/11)	**100%**
TGM3	rs214803	**0.98**	**A**	AALGVQSINWQkAFNR	**65.4%**	(34/52)	**100%**
LRRC15	rs13070515	**0.395**	**A**	ELSlGIFGPMPNLR	**4.8%**	(1/21)	**100%**
LRRC15	rs13070515	**0.944**	**G**	ELSPGIFGPMPNLR	**12.0%**	(6/50)	**100%**
KRT40	rs150812789	**0.433**	**G**	TASALEIELQAQQSLTESLECTVAETEAQYSSQLAQIQr / LIDNLENQLAEIR	**4.5%**	(1/22)	**100%**
KRT40	rs150812789	**0.931**	**A**	TASALEIELQAQQSLTESLECTVAETEAQYSSQLAQIQCLIDNLENQLAEIR	**0.0%**	(0/46)	**na**
KRT33A	rs12937519	**0.466**	**A**	QVVSSSEQLQSYQvEIIELR	**10.0%**	(3/30)	**100%**
KRT37	rs9910204	**0.475**	**A**	TSFYSTSSCPLcCTMAPGAR	**4.0%**	(1/25)	**100%**
KRT37	rs9910204	**0.934**	**C**	TSFYSTSSCPLGCTMAPGAR	**6.1%**	(3/49)	**100%**
KRT32	rs2071563	**0.485**	**A**	LEGEINmYR	**11.4%**	(4/35)	**100%**
KRT32	rs2071563	**0.832**	**G**	LEGEINTYR	**53.2%**	(25/47)	**89%**
GSTP1	rs1695	**0.544**	**G**	YvSLIYTNYEAGKDDYVK	**0.0%**	(0/36)	**na**
GSTP1	rs1695	**0.905**	**A**	YISLIYTNYEAGKDDYVK	**2.2%**	(1/46)	**100%**
KRT81	rs2071588	**0.56**	**G**	GLTGGFGSHSVCr	**100.0%**	(26/26)	**100%**
BLMH	rs1050565	**0.562**	**C**	HVPEEVLAVLEQEPIvLPAWDPMGALA-	**10.7%**	(3/28)	**100%**
BLMH	rs1050565	**0.884**	**T**	HVPEEVLAVLEQEPIILPAWDPMGALA-	**32.6%**	(15/46)	**100%**
KRT34	rs2239710	**0.575**	**T**	SQYEALVEINR	**73.5%**	(25/34)	**100%**
KRT83	rs2852464	**0.62**	**C**	DLNMDCmVAEIK	**36.8%**	(14/38)	**100%**
KRT83	rs2852464	**0.863**	**G**	DLNMDCIVAEIK	**84.4%**	(38/45)	**100%**
KRT35	rs2071601	**0.708**	**C**	TNCSaRPICVPCPGGRF-	**8.6%**	(3/35)	**100%**
KRT35	rs2071601	**0.798**	**G**	TNCSPRPICVPCPGGRF-	**7.1%**	(3/42)	**100%**
KRT35	rs743686	**0.778**	**G**	VSAMYSSSpCKLPSLSPVAR	**57.1%**	(20/35)	**100%**
KRT35	rs743686	**0.702**	**A**	VSAMYSSSSCKLPSLSPVAR	**78.6%**	(22/41)	**100%**
GSDMA	rs7212938	**0.731**	**G**	ALETVQER	**12.5%**	(5/40)	**100%**
GSDMA	rs7212938	**0.781**	**T**	ALETlQER	**2.9%**	(1/34)	**100%**

Table 4. Specificity and Sensitivity of Peptides Containing nsSNPs.

Peptides containing changes in primary structure due to non-synonymous single nucleotide polymorphisms (rs#; nsSNPs), were used to predict the genotype (nuc) of European American subjects as indicated in **Table 3**. The sensitivity of each peptide as a marker of genotype (assigned genotype/total genotypes) and specificity (1-False Positive Rate (%)) are indicated. Peptides (peptide) are ranked in order of the frequency (*f*) with which the minor allele occurs in the European population. The number of detected and total genotypes is indicated in parentheses.

Discussion

We were able to identify 36 nsSNP-containing peptides within this study, 25 of which were confirmed directly using Sanger sequencing. Assuming the use of the product rule where gene boundaries coincided with boundaries of linkage disequilibrium, we were able to use these to calculate an average overall profile frequency of 1 in 368. There is a large range of final results however, from 1 in 1 (L1.016) to 1 in 9000 (L1.058). The major factor determining the outcome of these calculations is the number of unique peptides identified in the analysis (**Figure 2**). Improved methods of biochemical digestion and increased sensitivity and efficiency in mass spectrometry data acquisition will increase the level of unique peptide identification and as a result increase the power of discrimination values obtained in the analysis. Improved, specialized instrumentation configurations also have the potential to improve the powers of discrimination and likelihood scores of biogeographic background[9].

The correct application of the product rule is predicated on finding the pair-wise linkage disequilibrium analyses for nsSNP loci that are not in the haplotype databases. A reduced dependence on rare loci will also reduce the role of disequilibria in the analysis. Nevertheless we are now in a position to determine the appropriate boundaries beyond which the product rule can be assumed and allow the development of "haplotype" frequencies within blocks of linkage disequilibrium to be calculated.

The issue of confidence intervals was not addressed in this study. Geneticists tend to focus on absolute allelic frequencies. However, the confidence in the power of discrimination will be improved if more common nsSNP loci are identified and there is less dependence on rare loci.

Future Work

Continuation of this project will require a more thorough examination of the physical and biochemical treatment of hair and refinement of the acquisition protocols. Expansion of this technology and increasing applicability will require analysis of additional subjects and other protein sample types, such as bone, teeth, and other tissues. Increased diversity of samples from other biogeographic backgrounds will also increase the applicability of the method both for the purposes of associating a given protein sample with an individual and making conclusions about its biogeographic background. The methodology developed using common nsSNPs can also be readily applied to potential work on private nsSNPs or nsSNPs with a tight biogeographic distribution.

References

1. Committee on Identifying the Needs of the Forensic Sciences Community, N.R.C., *Strengthening Forensic Science in the United States: A Path Forward*, D.o. Justice, Editor. 2009.
2. Butler, J.M., *Fundamentals of Forensic DNA Typing*. 2010: Academic Press.
3. Smith, C.I., et al., *The thermal history of human fossils and the likelihood of successful DNA amplification.* J Hum Evol, 2003. **45**(3): p. 203-17.
4. Ottoni, C., et al., *Preservation of ancient DNA in thermally damaged archaeological bone.* Naturwissenschaften, 2009. **96**(2): p. 267-78.
5. Callaway, E., *Proteins help solve taxonomy riddle.* Nature, 2013. **503**(7474): p. 18-9.
6. Misner, L.M., et al., *The correlation between skeletal weathering and DNA quality and quantity.* J Forensic Sci, 2009. **54**(4): p. 822-8.
7. Michalski, A., J. Cox, and M. Mann, *More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.* J Proteome Res, 2011. **10**(4): p. 1785-93.
8. Altelaar, A.M. and A.J. Heck, *Trends in ultrasensitive proteomics.* Curr Opin Chem Biol, 2012.
9. Hebert, A.S., et al., *The One Hour Yeast Proteome.* Mol Cell Proteomics, 2013.
10. Budowle, B. and A. van Daal, *Forensically relevant SNP classes.* Biotechniques, 2008. **44**(5): p. 603-8, 610.
11. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
12. van Oorschot, R.A., K.N. Ballantyne, and R.J. Mitchell, *Forensic trace DNA: a review.* Investig Genet, 2010. **1**(1): p. 14.
13. Bertrand, L., et al., *Microbeam synchrotron imaging of hairs from ancient Egyptian mummies.* J Synchrotron Radiat, 2003. **10**(Pt 5): p. 387-92.
14. Thibaut, S., et al., *Transglutaminase-3 enzyme: a putative actor in human hair shaft scaffolding?* J Invest Dermatol, 2009. **129**(2): p. 449-59.
15. Rice, R.H., *Proteomic analysis of hair shaft and nail plate.* J Cosmet Sci, 2011. **62**(2): p. 229-36.
16. Lee, Y.J., R.H. Rice, and Y.M. Lee, *Proteome analysis of human hair shaft: from protein identification to posttranslational modification.* Mol Cell Proteomics, 2006. **5**(5): p. 789-800.
17. Rogers, M.A., et al., *The human type I keratin gene family: characterization of new hair follicle specific members and evaluation of the chromosome 17q21.2 gene domain.* Differentiation, 2004. **72**(9-10): p. 527-40.
18. Rogers, M.A., et al., *Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13.* J Invest Dermatol, 2005. **124**(3): p. 536-44.

Appendices

Appendix 1. Biogeographic Background of Subjects using 129 racially informative SNPs

Appendix 2. SOP: Hair Trypsinization Protocol

Appendix 3. LC/MSMS Data Acquisition

Appendix 4. Primer Sequences Used for Genotyping with Sanger Sequencing

Appendix 5. Genotyping based on Sanger Sequencing

Appendix 6. Pair-Wise Linkage Disequilibrium Analysis

Appendix 7. Development of PBIT FASTA Database File

Appendix 8. Peptide Selection Method

Appendix 9. Overall Profile Frequency Recalculated with Strictest Product Rule Application

Appendix 10. Primer / Factors for Consideration

Appendix 1. Biogeographic Background of Subjects using 129 racially informative SNPs

Protein Study Ancestry Estimations

SampleID	% EUR	% EAS	% ISUB	% IDA	% SSAF	SD EUR	SD EAS	SD ISUB	SD IDA	SD SSAF
00642-001	100	0	0	0	0	3	0	0	0	0
00642-002	100	0	0	0	0	3	0	0	0	0
00642-003	100	0	0	0	0	3	0	0	0	0
00642-004	100	0	0	0	0	3	0	0	0	0
00642-005	100	0	0	0	0	3	0	0	0	0
00642-006	100	0	0	0	0	3	0	0	0	0
00642-007	100	0	0	0	0	3	0	0	0	0
00642-008	100	0	0	0	0	3	0	0	0	0
00642-009	100	0	0	0	0	3	0	0	0	0
00642-010	77	0	23	0	0	8	0	8	0	0
00642-011	91	0	0	0	9	3	0	0	0	3
00642-012	100	0	0	0	0	3	0	0	0	0
00642-013	100	0	0	0	0	3	0	0	0	0
00642-014	100	0	0	0	0	3	0	0	0	0
00642-015	100	0	0	0	0	3	0	0	0	0
00642-016	100	0	0	0	0	3	0	0	0	0
00642-017	100	0	0	0	0	3	0	0	0	0
00642-018	54	0	20	26	0	5	0	5	2	0
00642-019	100	0	0	0	0	3	0	0	0	0
00642-020	100	0	0	0	0	3	0	0	0	0
00642-021	100	0	0	0	0	3	0	0	0	0
00642-022	78	0	0	22	0	2	0	0	2	0
00642-023	100	0	0	0	0	3	0	0	0	0
00642-024	69	0	19	12	0	8	0	9	2	0
00642-025	78	0	22	0	0	6	0	6	0	0
00642-026	100	0	0	0	0	3	0	0	0	0
00642-027	88	0	0	12	0	2	0	0	2	0
00642-028	100	0	0	0	0	3	0	0	0	0
00642-029	100	0	0	0	0	3	0	0	0	0
00642-030	100	0	0	0	0	3	0	0	0	0
00642-031	100	0	0	0	0	3	0	0	0	0
00642-032	100	0	0	0	0	3	0	0	0	0
00642-033	100	0	0	0	0	3	0	0	0	0
00642-034	88	0	0	0	12	2	0	0	0	2
00642-035	100	0	0	0	0	3	0	0	0	0
00642-036	100	0	0	0	0	3	0	0	0	0
00642-037	100	0	0	0	0	3	0	0	0	0
00642-038	100	0	0	0	0	3	0	0	0	0
00642-039	100	0	0	0	0	3	0	0	0	0
00642-040	100	0	0	0	0	3	0	0	0	0
00642-041	100	0	0	0	0	3	0	0	0	0
00642-042	100	0	0	0	0	3	0	0	0	0
00642-043	90	0	0	0	10	2	0	0	0	2
00642-044	100	0	0	0	0	3	0	0	0	0
00642-045	100	0	0	0	0	3	0	0	0	0
00642-046	100	0	0	0	0	3	0	0	0	0
00642-047	100	0	0	0	0	3	0	0	0	0
00642-048	100	0	0	0	0	3	0	0	0	0
00642-049	100	0	0	0	0	3	0	0	0	0
00642-050	100	0	0	0	0	3	0	0	0	0
00642-051	100	0	0	0	0	3	0	0	0	0
00642-052	100	0	0	0	0	3	0	0	0	0
00642-053	100	0	0	0	0	3	0	0	0	0
00642-054	100	0	0	0	0	3	0	0	0	0
00642-055	100	0	0	0	0	3	0	0	0	0
00642-056	100	0	0	0	0	3	0	0	0	0
00642-057	100	0	0	0	0	3	0	0	0	0
00642-058	100	0	0	0	0	3	0	0	0	0
00642-059	100	0	0	0	0	3	0	0	0	0
00642-060	100	0	0	0	0	3	0	0	0	0

EAS = East Asian
ISUB = Indian Subcontinent
IDA = Indigenous American
SSAF = Sub-Saharan African

Before hair samples were processed, DNA from each subject was evaluated for biogeographic background using 129 racially informative SNPs. All subject self-identified as European (EUR) however some individuals were determined to have some admixture from other genetic groups. In order to maintain the integrity of the study subjects 10, 11, 18, 22, 24, 25, 27, 34, and 43 were excluded from further treatment and analysis.

Appendix 2. SOP: Hair Trypsinization Protocol

Effective Date: 20 November 2013 11/30/13 4:16 PM

Document ID number (revision): 003-1

Approval History: G. Parker 11/20/13

Purpose: To maximize the extraction of peptides from hair and prepare the sample for application to tandem liquid chromatography Mass Spectrometry

Protocol:

Equipment:

OMNI Bead Ruptor 24, model: 19-010, OMNI-International Inc.
Vortex Bench top Centrifuge: 15,000g
Rotator: MACSmix Miltenyi Biotec Inc. model# 2019
Analytical Balance. Mettler (0.1 mg sensitivity)

Materials:

(All non-organic reagents cleaned by passage through Waters SepPak Ct18)

Stocks:
Ammonium Bicarbonate (1 M)
DTT stock (1 M)
Iodoacetamide (1M)
Trypsin-TPCK (Worthington Enzymes Inc. cat# TRSEQZ)
Protease-Max (Promega cat# V207A)
Acetonitrile
Methanol
Ethanol
Water (MilliQ, MΩ>18.3)
"LoBind" 1.5 ml microcentrifuge tubes (Eppendorf cat# 022431081)
Rein. Ceramic Bead Kit, 2.8 mm (Omni-International # 19-628)

Solutions:
20% Methanol
50% Ethanol
Extraction Buffer:
- 1.5 M Urea
- 0.125 M DTT
- 0.1 M ammonium bicarbonate
- 0.01% (w/v) Protease-Max

Digestion Mixture:
- 0.08 M DTT
- 0.04 M ammonium bicarbonate
- 0.0125% (w/v) Protease-Max
- 0.5 µg/µL Trypsin-TPCK

Hair sample (10.5 ± 0.5 mg)

Method:

1. Prepare work surface (glass) by thorough cleaning with water, 20% methanol and acetonitrile. Use fresh gloves and Kimwipes.
2. Number and weigh each ceramic bead containing vial.
3. Place 10.5 mg of Hair into vials containing ceramic beads, record final weight. (variation = ± 0.5 mg) Samples can be stored at RT (20 to 23ºC).
4. Add 200µl of Extraction Buffer (1.5 M Urea, 0.125 M DTT, 0.1 M ammonium bicarbonate, 0.01% (w/v) Protease-Max).
5. Mill for 3 minutes at 4.5 m/s
6. Mix gently overnight (16 to 24h), on Rotator
7. Add 50 µL of 1M iodoacetamide
8. Mill for 3 minutes at 4.5 m/s
9. Rotate in Darkened room for 30 minutes
10. Add 200 µl of Digestion Mixture (0.08 M DTT, 0.04 M ammonium bicarbonate, 0.0125% (w/v) Protease-Max, 0.5 µg/µL Trypsin-TPCK)
11. Rotate overnight at RT (16 to 24h)
12. Centrifuge vials at 10,000g x 15 minutes. Aliquot 400 µL of supernatant into a 1.5 ml "LoBind" microcentrifuge tube.
13. Centrifuge tubes at 10,000 g x 15 minutes.
14. Aliquot SN into Agilent Sample Vials with glass insert.
15. Apply to MS.

Appendix 3. LC/MSMS Data Acquisition

Successful use of hair proteins for identification of individuals requires that large numbers of unique peptides be identified in each sample. Hair samples were processed by mechanical milling, chemical treatment (reduction and acylation), and tryptic digestion. The processed hair samples were then analyzed by liquid chromatography-tandem mass spectrometry (LC/MSMS). The goals of the LC/MSMS analysis are first to identify the protein by comparing measured amino acid sequences in the peptide to reference databases and then to identify single amino acid polymorphisms in those peptides. High quality LC/MSMS analysis to achieve these goals requires both optimized separation by LC and optimized MSMS.

The LC separation was achieved using gradient elution on a C18 column. Separation conditions were optimized by analyzing a standard hair sample (pooled processed hair from participant U3) and maximizing the number of unique peptides identified from the sample. This pooled sample was also used as a quality-control sample during optimized sample analysis runs to confirm that the LC/MSMS instrument was operating correctly. Elution gradient composition and time, column temperature, and injection volume were varied to determine optimal conditions for these parameters. The optimized separation conditions and instrument details for the LC separation are listed in the following table.

Instrument Configuration	Agilent 1290 Infinity Autosampler Agilent 1290 Infinity Pump Agilent 1290 Infinity Column Compartment
Separation Column	Agilent AdvanceBio Peptide Map PN 655750-902 2.1 mm x 100 mm C18, 2.7 micron particles, 120Å pore size
Guard Column	Agilent AdvanceBio Peptide Map Guard PN 851725-911 2.1 mm x 5 mm C18, 2.7 micron particles, 120Å pore size
Column Temperature	50 °C
Injection Volume	20 microliters
Mobile Phase A	Water, 0.1% formic acid
Mobile Phase B	Acetonitrile, 0.1% formic acid
Flow rate	200 microliter/min
Binary Gradient Profile	5% B for 5 minutes (diverted to waste) Linear ramp to 50% B in 120 minutes Step to 95% B Hold at 95% B for 5 minutes
Post-separation column clean-up and re-equilibration	Cycle 4 times: 1 min linear ramp to 5% B followed by 1 min linear ramp to 95% B 1 min linear ramp to 5% B Hold at 5% B for 15 minutes
Injection needle wash (external)	5 sec flush with 1:1 water/isopropanol

As the peptides elute from the separation column they are introduced into the mass spectrometer using electrospray ionization and analyzed using auto MSMS. In the auto MSMS approach, the instrument repeats an analysis cycle that starts with an MS scan followed by a prescribed number of MSMS scans using precursor ions identified in the initial MS scan. The MS scan analyzes all of the intact peptide ions (no collision-induced dissociation) and applies a set of rules (based on intensity, charge state, and inclusion and exclusion lists) to tabulate the precursor ions that will be fragmented using collision-induced dissociation in the subsequent series of MSMS analyses in the cycle. In each MSMS step, a single precursor ion is selected and is collided with N_2 gas at a defined collision energy. Ideally, this collision-induced-dissociation step produces series of fragments where the peptides are cleaved at each peptide linkage and the fragment ions reveal the amino acid sequence in the precursor ion.

MSMS conditions were optimized using the same standard hair sample (pooled processed hair from participant U3) that was used for LC optimization. Auto MSMS parameters were chosen that maximized the number of unique peptides identified in the standard hair sample. The major parameters for the QTOF are listed in the following table.

Instrument Configuration	Agilent 6530 Accurate-Mass Q-TOF LC/MS Dual Agilent Jet Stream ESI Ion Source
Instrument Mode	Mass range: 3200 m/z Extended Dynamic Range (2 GHz)
MSMS parameters	
MS Min Range (m/z)	100
MS Max Range (m/z)	3000
MS Scan Rate (spectra/sec)	6.00
MS/MS Scan Rate (spectra/sec)	1.50
Isolation Width MS/MS	Medium (~4 amu)
Collision Energy Equation	
Charge	All
Slope	3
Offset	2
Precursor Selection	
Max Precursors Per Cycle	5
Threshold (Absolute)	1000
Threshold (Relative %)	0.010
Precursor abundance based scan speed	Yes
Target (counts/spectrum)	25000.000
Use MS/MS accumulation time limit	Yes

Use dynamic precursor rejection	No
Purity Stringency (%)	75.000
Purity Cutoff (%)	30.000
Isotope Model	Peptides
Active exclusion enabled	Yes
Active exclusion excluded after (spectra)	6
Active exclusion released after (min)	0.50
Sort precursors	By charge state then abundance
Static Exclusion Ranges	
Start (m/z)	100
End (m/z)	300
Charge State Preference (in order)	2, 3, 1, >3
Source Parameters	
Gas Temp (°C)	320
Gas Flow (l/min)	8
Nebulizer (psig)	27
Sheath Gas Temperature	380
Sheath Gas Flow	12
V_{Cap}	3750
Nozzle Voltage (V)	500
Fragmentor	150
Skimmer1	65
Octopole RF (Peak)	750
Reference Masses	121.05090
	922.00980

Processed hair samples were run in batches. Before each batch, the mass spectrometer was tuned and calibrated according to the manufacturer's instructions. To avoid cross contamination, two blanks were blanks were run at the start of the batch and one was run before each sample. The blanks were run with a faster gradient (ramp 5% to 50% in 20 minutes rather than 120 minutes) than the sample runs, but the rest of the elution program (including the column clean-up steps) was identical to the sample runs. Blank runs were used to confirm that this clean-up protocol was sufficient to remove residual peptides from the separation column. The quality control sample was run at the start of the batch and after every five samples. Occasionally (at least once per batch of samples) a blank was run in place of a sample to confirm that no sample carry-over was affecting the data.

Appendix 4. Primer Sequences Used for Genotyping with Sanger Sequencing

	rs Number	Hg19 Coordinates for variant	Annealing temp	Forward Primer Name	Forward Primer Sequence
	rs6580873	chr12:52681925	64	PPHP1F	CTGCGGGTGACAATGTC
	rs2852464	chr12:52710721	64	PPHP2F	GTGTCCTGTGCCGCTCAC
	rs6503627	chr17:39553547	64	PPHP3F	TTCTGCTGGAGCTCCTCA
	rs13070515	chr3:194080916	61.4	PPHP6F	GGGAGATGAAGCTGATC
	rs1050565	chr17:28576076	58.3	PPHP8F	TTCAGTCCCTGGATCTGT
	rs371749	chrX:118603844	58	PPHP9F	TGGCTTCCTTCCTGTCTG
	rs10805890	chr5:73992881	54	PPHP12F	TCATTGAGTTCAAGACA
	rs12937519	chr17:39503163	64	PPHP15F	GGTGGGGAAGATTGAG
	rs1695	chr11:67352689	54	PPHP17F	TCCCCAGTGACTGTGTG
	rs1455555	chr18:61170782	54	PPHP18F	CCAATGCCAAGGTCAA
	rs3894194	chr17:38121993	58.3	PPHP19F	TCAGAGCCGGGTAAACT
	rs2239710	chr17:39535859	54.8	PPHP20F	TGAGTGGCCATGTGCTT
	rs143043662	chr17:39913771	63.4	PPHP23F	GCTTGAAGAGGGAGTTC
	rs2071588	chr12:52685096	63.4	PPHP24F	GACCGACACGGTGGTGA
	rs112544857	chr17:39551763	58	PPHP25F	AAGGGGAGAAAGGATC
	rs17678945	chr12:53070174	58	PPHP26F	GTCCCTTCCTCACCTGCT
	rs13060627	chr3:194080983	54.8	PPHP27F	CCCAAAGAGAGTAAGAC
	rs56030650	chr17:38131187	54.8	PPHP28F	ACATGAGGCCTCGAAAA
	rs2071563	chr17:39619115	54	PPHP29F	TTCCTATAACCTCCCAGA
	rs2071561	chr17:39622068	54.8	PPHP30F	GTTCCTCCCCCTCTGCTT
	rs6929069	chr6:7581636	64	PPHP31F	TGGCCACCTGAGGGAAA
	rs7212938	chr17:38122680	64	PPHP33F	CTGCAGACCCACACAGAC
	rs41283425	chr17:39925713	62.1	PPHP34F	GCCCCAGTCAGACTCATC
	rs743686	chr17:39637244	62.5	PPHP35F	TCGTTCAGGGATTGCATC
	rs150812789	chr17:39135207	62.1	PPHP36F	CATCAGCCTCTTTGTGAC
	rs951773	chr12:52774235	62.5	PPHP37F	GATGCCTAAGTCCTGACTT
	rs36022742	chr1:153520954	62	PPHP38F	CCCCAGGTCTCTCTTTGC
	rs214803	chr20:2290333	64	PPHP39F	GAATTTGGCTCTGCCCTT
	rs12450621	chr17:39525750	64	PPHP40F	CTGGCAGTGTGGGTGCCC
	rs112554450	chr12:52758810	64	PPHP41F	CCACTGGATATCTGCTCT
	rs200307345	chr11:1629558	62.1	PPHP42F	CTTGCAGCAGCAGACAG

Appendix 5. Genotyping based on Sanger Sequencing

LabID	KRT81_PPHP1_rs6580873	KRT83_PPHP2_rs2852464	KRT31_PPHP3_rs6503627	LRRC15_PPHP6_rs13070515	BLMH_PPHP8_rs1050565	SLC25A5_PPHP9_rs371749	HEXB_PPHP12_rs10805890	KRT33A_PPHP15_rs12937519	GSTP1_PPHP17_rs1695	SERPINB5_PPHP18_rs1455555	GSDMA_PPHP19_rs3894194	KRT34_PPHP20_rs2239710	JUP_PPHP23_rs143043662	KRT81_PPHP24_rs2071588	KRT31_PPHP25_rs112544857	KRT1_PPHP26_rs17678945	LRRC15_PPHP27_rs13060627	GSDMA_PPHP28_rs56030650	KRT32_PPHP29_rs2071563	KRT32_PPHP30_rs2071561	DSP_PPHP31_RS6929069	GSDMA_PPHP33_RS7212938	JUP_PPHP34_RS41283425	KRT35_PPHP35_RS743686	KRT40_PPHP36_Rs150812789	KRT84_PPHP37_RS951773	S100A3_PPHP38_RS36022742	TGM3_PPHP39_RS214803	KRT33B_PPHP40_Rs12450621	KRT85_PPHP41_RS112554450	KRTAP5-3_PPHP42_RS200307345	KRT32_PPHP43_rs11078993	S100A3_PPHP44_rs41265164	KRT35_PPHP45_rs2071601	KRT35_PPHP45_rs12451652	KRT85_PPHP46_rs61630004	KRT37_PPHP47_rs9910204	KRT32_PPHP48_rs72830046
LLNL01	C/C	C/C	G/G	G/G	T/T	G/G	A/A	G/G	A/A	A/A	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/A	G/T	G/G	G/T	C/T	G/A	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/T
LLNL02	C/C	G/C	G/G	A/A	T/T	G/T	A/A	G/G	A/A	A/A	G/G	C/C	C/C	C/C	G/G	C/C	T/T	C/C	G/A	G/T	G/G	T/T	C/T	G/G	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/C	A/A	C/T
LLNL03	C/C	C/C	A/A	G/G	T/C	G/T	A/G	G/A	A/G	G/G	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/A	G/G	G/A	G/G	C/C	A/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL04	C/C	C/C	A/A	A/A	T/T	T/T	A/A	G/A	A/G	G/G	G/G	C/T	C/C	C/C	G/G	C/C	T/T	A/A	G/A	G/G	G/G	T/T	C/C	A/A	A/A	C/C	C/C	C/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL05	A/C	G/C	A/A	G/G	C/C	G/G	A/A	G/A	A/A	A/G	G/G	C/C	C/C	C/C	G/G	C/C	C/C	C/C	G/G	G/T	G/G	G/T	C/C	G/A	A/A	C/C	C/C	A/A	C/C	C/T	Failed	C/C	G/G	C/G	C/C	C/C	C/A	C/T
LLNL06	A/A	G/G	G/G	G/G	C/C	G/G	A/A	G/A	A/G	A/G	G/G	C/T	C/C	C/C	G/A	C/C	C/C	C/C	G/A	G/T	G/G	T/T	C/C	A/A	A/A	T/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/T	C/C	C/C	C/C
LLNL07	C/C	G/G	G/G	G/G	T/C	G/T	A/A	G/A	A/A	A/A	C/T	C/C	G/G	G/G	C/C	C/C	C/A	G/A	G/T	G/G	G/G	C/C	G/A	A/A	T/T	C/C	C/C	C/C	C/C	C/C	A/A	C/C	G/G	C/G	C/T	C/C	C/C	C/C
LLNL08	A/C	G/G	G/G	G/G	T/C	G/G	A/G	G/A	A/G	G/G	G/A	C/T	C/C	C/G	G/G	C/C	C/C	C/A	G/A	G/T	G/A	G/T	C/C	G/G	G/G	T/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/T
LLNL09	C/C	G/G	G/G	G/G	T/T	G/G	A/A	G/A	A/G	A/G	A/A	C/T	C/C	G/G	G/G	C/C	C/C	A/A	G/G	T/T	G/G	G/G	C/C	G/A	G/A	C/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	C/T	C/C	C/A	C/T
LLNL12	A/C	G/C	G/G	G/A	T/C	G/G	A/G	G/G	G/A	G/A	C/C	C/C	C/C	C/C	G/G	C/C	C/T	C/A	A/A	G/G	G/G	G/T	C/C	A/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL13	C/C	G/C	G/G	G/G	C/C	G/G	A/A	G/G	A/G	A/G	G/G	C/C	C/C	C/C	G/G	C/C	C/C	C/C	G/G	G/G	G/A	T/T	C/C	G/G	G/G	C/C	C/C	C/C	C/C	C/T	A/A	C/C	G/G	C/C	C/C	C/C	A/A	T/T
LLNL14	C/C	G/C	G/G	G/A	T/T	G/T	A/G	G/G	A/A	A/G	A/A	C/C	C/C	C/C	G/G	C/C	C/T	A/A	G/G	T/T	G/G	G/G	C/C	G/G	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/C	A/A	T/T
LLNL15	C/C	G/G	G/G	G/G	T/C	G/G	A/A	G/G	A/A	G/A	C/T	C/C	G/G	G/G	C/C	C/C	C/A	G/A	G/G	G/G	G/T	C/C	G/A	G/A	C/T	C/C	A/A	C/C	C/C	A/A	C/C	C/C	G/G	C/G	C/C	C/C	C/A	C/T
LLNL16	C/C	G/C	G/G	G/G	C/C	G/G	A/A	G/G	A/G	G/G	G/G	C/C	C/C	C/C	G/G	C/C	C/C	C/C	G/A	G/G	G/G	T/T	C/T	G/G	G/G	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/C
LLNL17	C/C	G/C	G/G	G/G	T/T	G/T	A/G	G/G	A/G	A/G	A/A	C/C	C/C	C/C	G/G	C/C	C/C	A/A	G/G	T/T	G/G	G/G	C/C	G/G	G/A	C/C	C/C	C/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/C	C/A	T/T
LLNL19	C/C	G/G	G/G	G/G	T/T	G/G	A/A	G/G	G/G	A/G	G/A	C/C	C/C	C/C	G/G	C/A	C/C	C/A	G/A	G/T	G/A	G/G	C/C	G/G	A/A	C/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/T
LLNL20	C/C	G/C	G/G	G/G	T/C	G/G	A/G	G/G	A/A	A/A	G/G	C/T	C/C	C/G	G/G	C/C	C/C	C/C	A/A	G/G	G/G	T/T	C/C	A/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL21	C/C	G/G	G/G	G/G	T/C	G/T	A/G	G/G	A/G	G/G	G/G	C/T	C/C	C/G	G/G	C/C	C/C	C/C	G/A	G/T	G/G	T/T	C/C	G/A	A/A	C/C	C/T	A/A	C/C	C/T	A/A	C/C	G/A	C/G	C/C	C/C	C/A	C/T
LLNL23	C/C	G/C	G/G	G/G	T/T	G/G	A/G	G/A	A/G	A/A	C/T	C/C	C/G	G/G	C/C	C/C	A/A	G/A	G/T	G/A	G/G	C/C	G/A	G/A	C/C	C/C	A/A	C/C	C/C	Failed	C/C	G/G	C/G	C/T	C/C	C/C	C/C	C/C
LLNL26	C/C	G/G	G/A	G/A	T/C	G/G	A/A	G/A	G/G	A/A	G/G	C/C	C/C	C/G	G/A	C/C	C/T	C/C	G/G	G/T	G/A	T/T	C/C	G/A	A/A	C/T	C/C	A/A	C/C	C/C	A/A	C/T	G/G	C/G	C/C	C/C	C/C	C/C
LLNL28	C/C	G/C	G/G	G/A	T/C	G/G	A/A	G/G	A/A	A/G	A/A	C/T	C/C	C/C	G/G	C/C	C/T	A/A	A/A	G/G	G/G	G/G	C/C	G/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	C/C	C/C	C/C	C/C
LLNL29	A/C	G/G	A/A	G/A	T/C	G/G	A/G	G/A	A/A	A/A	C/T	C/C	C/C	G/G	C/C	C/T	C/A	G/A	G/G	G/G	C/C	A/A	A/A	C/T	C/C	A/A	C/C	C/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL30	C/C	G/C	G/G	G/A	T/T	G/G	A/A	G/G	A/A	A/A	A/A	C/C	C/C	C/C	G/G	C/C	C/T	A/A	G/A	G/T	G/G	G/G	C/C	G/G	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/T	C/A	C/T
LLNL31	C/C	C/C	A/A	G/G	T/T	G/T	A/A	G/A	A/G	G/G	G/G	C/C	C/C	C/C	G/G	C/C	C/C	C/C	G/G	G/G	G/G	T/T	C/C	G/A	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	C/C	C/C	C/A	C/T
LLNL32	C/C	G/C	G/A	G/G	T/T	G/T	A/A	A/A	A/G	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/A	G/G	G/G	G/T	C/C	A/A	A/A	C/C	C/C	A/A	C/T	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL33	C/C	G/G	A/A	G/G	T/T	G/G	A/G	G/A	A/G	A/G	A/A	C/T	C/C	G/G	G/G	C/A	C/C	C/A	G/A	G/G	G/G	G/G	C/C	G/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL35	C/C	C/C	G/G	G/A	T/T	T/T	A/A	G/G	G/G	A/G	A/A	C/T	C/C	C/C	G/G	C/C	C/T	C/C	A/A	G/G	G/G	G/T	C/C	A/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	failed	G/G	C/C	C/T	C/C	C/C
LLNL36	C/C	G/C	G/G	G/G	T/C	T/T	A/G	G/G	A/G	G/G	C/C	C/C	C/C	G/G	C/C	C/C	C/A	G/G	T/T	G/G	G/T	C/C	G/G	G/G	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/C	A/A	T/T
LLNL37	C/C	G/C	G/G	G/G	T/C	G/G	A/A	G/A	A/A	A/A	C/T	C/C	C/C	G/G	C/C	C/C	A/A	G/A	G/T	G/G	G/G	C/C	A/A	G/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/T	C/C	C/C	C/C	C/C
LLNL38	C/C	G/C	G/G	G/A	C/C	T/T	A/A	G/G	A/A	A/G	G/A	C/C	C/C	C/C	G/G	C/C	C/T	C/A	G/A	G/T	G/A	G/T	C/C	G/G	G/G	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/T
LLNL39	C/C	G/C	A/A	G/A	T/T	G/G	A/A	A/A	A/A	G/A	C/C	C/C	C/C	G/G	C/C	C/T	C/A	G/A	G/G	A/A	G/T	C/C	G/A	A/A	C/C	C/T	A/A	C/C	C/C	A/A	C/C	G/A	C/G	C/C	C/C	C/C	C/C	C/C
LLNL40	A/C	G/C	G/G	G/G	T/C	G/T	A/A	G/G	A/A	A/A	G/A	T/T	C/C	C/C	G/G	C/C	C/C	C/A	A/A	G/G	G/G	G/T	C/C	A/A	A/A	C/T	C/T	A/A	C/C	C/C	A/A	C/C	G/A	G/G	C/C	C/C	C/C	C/C
LLNL41	A/C	G/G	G/G	G/G	T/T	G/T	A/G	G/G	A/G	A/G	G/A	T/T	C/C	C/C	G/G	C/C	C/C	C/A	A/A	G/G	G/G	G/T	C/C	A/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL42	C/C	G/C	G/A	G/G	T/C	G/G	A/G	A/A	A/G	G/A	C/T	C/C	C/G	G/G	C/C	C/C	C/A	G/G	G/T	G/G	G/T	C/C	A/A	G/A	C/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	G/G	C/T	C/C	C/C	C/C	C/C
LLNL44	C/C	G/C	G/G	G/G	T/C	G/T	A/A	G/G	A/G	A/A	A/A	C/C	C/C	G/G	G/G	C/A	C/C	A/A	A/A	G/G	G/G	G/G	C/C	G/A	A/A	C/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/A	C/C
LLNL45	A/A	G/C	G/G	G/A	T/T	T/T	A/A	G/G	G/G	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/T	C/A	G/A	G/T	G/G	G/T	C/T	G/A	A/A	C/C	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	C/C	C/C	C/C	C/T
LLNL46	C/C	G/C	A/A	G/A	T/T	T/T	A/A	G/A	A/G	A/G	A/A	C/C	C/C	C/G	G/G	C/C	C/T	C/A	G/G	G/T	G/G	G/G	C/T	G/A	A/A	C/C	C/C	C/A	C/C	C/C	A/A	C/C	G/G	C/G	C/C	C/A	C/T
LLNL47	A/C	G/G	G/G	G/A	T/T	G/G	A/A	G/A	A/G	A/G	G/A	T/T	C/C	C/G	G/G	C/C	C/T	C/A	G/A	G/T	G/G	G/T	C/C	A/A	G/A	C/C	C/C	C/A	C/C	C/C	A/A	C/C	G/G	G/G	C/T	C/C	C/C	C/C
LLNL48	C/C	G/G	G/G	G/G	T/T	G/T	A/A	G/A	A/G	A/G	G/G	C/T	C/C	G/G	G/G	C/A	C/C	C/A	G/A	G/T	G/G	G/T	C/C	G/A	A/A	T/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	C/T	C/C	C/C	C/C
LLNL49	C/C	C/C	G/G	G/A	T/C	G/G	A/A	A/A	A/A	G/G	G/A	T/T	C/C	C/C	G/G	C/C	C/T	C/A	G/G	T/T	G/G	G/G	C/C	A/A	A/A	C/C	C/T	A/A	C/C	C/C	A/A	C/C	G/A	G/G	T/T	C/C	C/C	C/C
LLNL50	C/C	G/C	A/A	A/A	T/T	G/G	A/A	G/A	A/G	A/A	A/A	C/C	C/C	C/C	G/G	C/A	T/T	A/A	G/G	G/T	G/G	G/G	C/C	G/A	G/A	C/C	C/C	A/A	C/C	C/C	A/A	failed	G/G	C/G	C/C	C/C	C/A	C/T
LLNL51	A/C	G/C	G/G	G/G	T/T	G/G	A/G	G/G	A/G	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	A/A	G/G	G/A	G/G	C/C	G/A	A/A	C/T	C/C	C/A	C/C	C/C	A/A	C/C	G/G	C/G	C/C	C/C	C/C	C/C
LLNL52	C/C	C/C	G/G	G/G	T/T	G/T	A/A	A/A	A/G	A/G	G/G	T/T	C/C	C/C	G/G	C/C	C/C	C/C	G/A	G/T	G/G	T/T	C/C	G/A	A/A	C/C	C/C	A/A	C/T	C/C	Failed	C/C	G/G	C/G	C/T	C/C	C/C	C/C
LLNL53	C/C	C/C	A/A	G/G	T/T	G/G	A/A	G/A	A/A	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/A	G/G	G/G	G/T	C/C	G/A	A/A	C/C	C/C	A/A	C/C	C/C	Failed	C/C	G/G	G/G	C/C	C/C	C/C	C/C
LLNL54	C/C	G/C	G/G	G/A	T/T	T/T	A/A	G/A	G/G	A/G	G/A	C/T	C/C	C/G	G/G	C/C	C/T	C/A	G/G	T/T	G/G	G/T	C/C	G/A	G/A	C/T	C/T	C/A	C/C	C/C	A/A	C/C	G/A	C/G	C/T	C/C	C/A	C/T
LLNL55	C/C	G/C	A/A	G/G	C/C	G/G	A/A	G/A	A/A	A/G	G/C	C/C	C/C	C/G	G/G	C/C	C/C	C/C	G/A	G/G	G/A	T/T	C/C	G/A	A/A	C/T	C/C	A/A	C/C	C/C	Failed	C/C	G/G	C/G	C/C	C/C	C/C	C/T
LLNL56	C/C	G/G	G/G	G/G	T/T	G/G	A/A	G/G	A/G	G/G	G/A	C/C	C/C	G/G	G/G	C/C	C/C	C/C	G/G	G/T	G/G	G/T	C/C	G/G	G/G	C/T	C/C	C/A	C/C	C/C	A/A	C/C	G/G	C/C	C/C	C/C	A/A	T/T
LLNL57	C/C	C/C	G/G	G/G	C/C	G/G	A/G	G/A	A/A	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/G	G/T	G/A	G/T	C/C	G/A	G/A	C/C	C/T	A/A	C/C	C/C	A/A	C/C	G/A	C/G	C/T	C/T	C/A	C/T
LLNL58	A/C	G/C	G/G	G/A	T/T	G/G	A/G	G/A	A/A	A/G	G/A	C/T	C/C	C/C	G/G	C/C	C/T	C/A	G/A	G/T	G/G	G/T	C/T	G/A	A/A	C/T	C/C	C/A	C/C	C/C	A/A	C/C	G/G	C/G	C/T	C/C	C/A	C/C
LLNL59	A/C	G/C	G/G	G/G	C/C	G/T	A/A	G/A	A/A	A/A	G/A	C/T	C/C	C/C	G/G	C/C	C/C	C/A	G/A	G/T	G/G	G/T	C/C	G/A	A/A	C/C	C/C	A/A	C/T	C/C	A/A	C/T	G/G	C/G	C/C	C/C	C/C	C/C
LLNL60	A/C	G/C	G/G	G/G	T/C	G/G	A/A	A/A	A/G	A/G	G/G	T/T	C/C	C/C	G/G	C/C	C/C	C/C	G/A	G/T	G/G	T/T	C/C	A/A	A/A	C/T	C/C	C/A	C/C	C/C	A/A	C/C	G/G	G/G	C/T	C/C	C/C	C/C
UVU03	C/C	G/G	G/G	G/A	T/C	G/G	A/G	G/G	A/G	G/G	C/T	C/C	G/G	G/G	C/A	C/T	C/C	A/A	G/G	G/G	T/T	C/C	G/A	G/A	C/C	C/T	A/A	C/C	C/C	A/A	C/C	G/G	C/C	C/G	C/C	C/C	A/G	C/C
UVU05	A/C	G/G	A/A	G/A	T/C	G/G	A/G	G/A	A/G	A/G	G/A	C/C	C/C	C/G	G/G	C/C	C/T	C/A	G/G	G/T	G/G	G/T	C/C	G/A	G/A	T/T	C/C	A/A	C/C	C/C	A/A	C/C	G/G	C/G	failed	C/C	C/A	C/T
UVU17	C/C	G/C	G/G	G/G	T/C	G/G	A/A	G/G	A/G	A/G	G/G	C/C	C/C	C/C	G/A	C/C	C/C	C/C	G/G	T/T	G/G	T/T	C/C	G/G	G/G	C/C	C/C	A/A	C/C	C/C	A/A	C/T	G/G	C/C	C/C	C/C	C/A	C/T

Genotyping of European Subjects Within Cohort

Validation of predicted DNA polymorphisms was carried out by first designing PCR primers, flanking the variant, using the Primer 3 program (Whitehead Institute for Biomedical Research). PCR reactions were carried out using the AccuPrime™ *Taq* DNA Polymerase System (Invitrogen™) following the manufacturer's specifications. PCR product was then treated with ExoSAP-IT® (Affymetrix) and subjected to Sanger Dideoxy Sequence analysis on an Applied Biosystems 3730xl 96-capillary DNA Analyzer by the DNA Sequencing Core Facility, University of Utah Health Science Cores. Polymorphisms in the DNA sequence were automatically called using CodonCode Corporation Aligner software and were subsequently viewed and individually verified by the user.

Appendix 6. Pair-Wise Linkage Disequilibrium Analysis

CHR 17 D' values

	MAF (EUR)	0.344	0.27	0.426	0.046	0.26	0.463	0.462	0.174	0.462
		KRT34	KRT37	***KRT32***	KRT32	KRT32	KRT32	KRT35	KRT35	***KRT35***
MAF (EUR)	**rs#**	**rs2239710**	**rs9910204**	**rs2071563**	**s1107899**	**s7283004**	**rs2071561**	**rs2071601**	**s1245165**	**rs743686**
0.344	***KRT34*** **rs2239710**		0.807	0.06	0.459	0.891	0.288	0.796	1	0.803
0.27	KRT37 **rs9910204**	0.807		0.767	0.06		0.755	0.933	1	0.932
0.426	***KRT32*** **rs2071563**	0.06	0.767		0.507	1	1	0.162	1	0.163
0.046	KRT32 **rs11078993**	0.459	0.06	0.507			1	1	1	1
0.26	KRT32 **rs72830046**	0.891		1			0.866	1	1	1
0.463	KRT32 **rs2071561**	0.288	0.755	1	1	0.866		0.281	1	0.25
0.462	KRT35 **rs2071601**	0.796	0.933	0.162	1	1	0.281		1	1
0.174	KRT35 **rs12451652**	1	1	1	1	1	1	1		1
0.462	***KRT35*** **rs743686**	0.803	0.932	0.163	1	1	0.25	1	1	

CHR 12 D' values

	MAF (EUR)	0.152	0.379	0.269	0.021
		KRT81	***KRT83***	KRT84	KRT1
MAF (EUR)		**rs6580873**	**rs2852464**	**rs951773**	**s17678945**
0.152	***KRT81*** **rs6580873**		0.689	0.058	1
0.379	***KRT83*** **rs2852464**	0.689		0.863	1
0.269	KRT84 **rs951773**	0.058	0.863		1
0.021	KRT1 **rs17678945**	1	1	1	

A pair-wise linkage disequilibrium analysis on nsSNPs identified in the proteomic datasets was conducted using the SNP Annotation and Proxy Search tool from the Broad Institute (http://www.broadinstitute.org/mpg/snap/ldsearchpw.php) (**Table 2.**). D' values were calculated with complete linkage (D' = 1) formatted in red, and no linkage (D' = 0) formatted in green. Loci (rs#) originating from within one gene are outlined. Loci corresponding to non-synonymous SNPs identified and in this study and confirmed with genotyping are indicated in bold and italics. Because pair-wise linkage disequilibrium analysis is affected by allele frequency, we also include the minor allele frequency for the European population in the 1000 Genomes project.

Appendix 7. Development of PBIT FASTA Database File

The PBIT database is a unique protein sequence database, developed for the express purpose of defining variant peptides that can then be detected for use in the identification of individuals. This database can be used in conjunction with any mass spectrometry analytical tool such as Xtandem, Sequest, Mascot, and SpectraST. The RefSeq protein database was used as a starting point for the PBIT protein reference database. The RefSeq protein sequence database human.protein.gpff.gz contains all known amino acid (aa) variant information, but is not in a format readily useful as a database for mass spectrometry software engines. From the UCSC ftp site ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/, the file snp137Common.txt.gz, which contains all of the common variants with frequencies >= 1%, can be downloaded using either of the following methods.

A) http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/
download snp137Common.txt.gz

or

B) http://genome.ucsc.edu/cgi-bin/hgTables
select drop-down menus as below
clade: Mammal, genome: Human, assembly: Feb. 2009 (GRCh37/hg19)
group: Variation and Repeats
table: snp137Common, track: Common SNPs(137)
region: genome
file type returned: gzip compressed
"get output" to download

The human.protein.gpff file contains reference sequences, but not necessarily unique sequences. First, 4817 duplicated sequences were removed from the database. Then for each sequence, the list of variants was gathered from two sources: the snp137Common.txt.gz file and the ESP 6500 db which contains SNPs, INDELs and coverage data for the ESP 6500 exomes (chromosomes 1-22, X, and Y). File ESP6500SI-V2-SSA137.dbSNP138-rsIDs.snps_indels.txt.tar.gz comes from the ftp site http://evs.gs.washington.edu/EVS/ at NHLBI. The snp137Common.txt file contains all of the common variants with MAF (minor allele frequency) >= 1%. The ESP6500 database contains data from various collaborators from 6503 samples for European American (EA) and African American (AA) individuals. The ESP 6500 database with 3.47 million variants, was filtered to pull all variants with either EA or AA MAF >= 0.5%.

All unique variants from these two sources were then used to create the variant sequences used in the PBIT database. Each reference sequence was duplicated once, and labeled the same as the reference sequence with the exception of the addition of the ".v1" string at the end of the NM number. The position of the variants in the sequence and their individual proximity was not a factor. If however, two or more variants occurred in the same position, the first variant in the list at that position was used. Stop variants were not used. The final PBIT database contains a reference sequence and variant sequence, if one or more variants exist in the sequence, for each protein sequence.

There are 34,383 NP_ loci, 1,833 XP_ loci, and 13 YP_ loci in human.protein.gpff for a total 36229 unique locus names for homo sapiens. The NM numbers are an identifier that differentiates between multiple assignments to the same gene, and are used in the PBIT database as a way of identifying sequence.

Large proteins presumably not involved in hair were removed from the file to facilitate run time. The list of proteins removed is as follows:
TTN, MUC16, OBSCN, NEB, MUC19, AHNAK, AHNAK2, MUC5B, MUC4, FCGBP, MUC12, LOC100289142, USH2A, MUC2, SSPO, HYDIN, RYR1.

The database is formatted in FASTA format.

SUMMARY OF PBIT FASTA database file:

A characterization of the database includes:
37% of proteins do not have variants > 0.5%
91% of proteins have 0 or 1 variants > 0.5%
99% of proteins have 0,1 or 2 variants > 0.5%

There are 36,229 protein sequences represented by a unique NM number.
350 protein sequences have at least one peptide with 3 variants
169 protein sequences have at least one peptide with 4 or more variants.

The largest number of variants per peptide is 27. It is in an extraordinarily large peptide in the MUC21 gene.

The number of variants in the ESP db with a MAF > 0.5% was 106,000 variants. The unique list of these variants is 67250. Of these, there are 31230 which were not identified in the snp137common.txt file. There are 13,585,949 rs numbers in the snp137Common.txt file and 31,230 NEW ESPdb rs numbers that have a frequency of > 0.5%, so the total number of rs numbers to use is 13,617,179.

The human.protein.gpff file which contains 36,229 genes and 732,776 variants was compared to each 13,617,179 variant in the combined ESPdb and snp137common file. Some of these variants are not necessarily associated with genes. The human.protein.gpff file matched 80,598 variants with the variants in the snp137common/ESPdb file. There were 9,491 genes with no variants. There were 19614 genes with a maximum of 1 variant in any peptide. There were 5772 genes with one or more peptides with 2 variants. There were 996 genes with one or more peptides with 3 variants. This left 356 genes with one or more peptides with 4 variants or more per peptide.

Even though there are 80,598 unique variants there are a total number of 127,099 variants in the database file, because of isoforms. Of these 127,099 variants 1518 are simple stop codons "*". 47 more are not simple stops e.g. "y*w". 6 variants are replaced by a '-' which is a deletion. These were not included in the variant database. 126,477 variants have a single amino acid replacement and 622 have multiple amino acid insertions.

The new variant database contains 53,476 sequences and the reference database contains 36,229 sequences

Of the 7,124 genes with two variants or more in a peptide(s), there were 1097 variants that share their variant position with only one other variant, 37 share with 2 other variants and 3 share their position with 3 other variants.

Appendix 8. Peptide Selection Method

Mass Spec data sets were collected and then converted to .mzXML format using MSConvert. The data sets were then analyzed using Trans-Proteomic Pipeline TANDEM and the database described above. The program Tandem2XML was used to convert the output .tandem data file to .xml format.

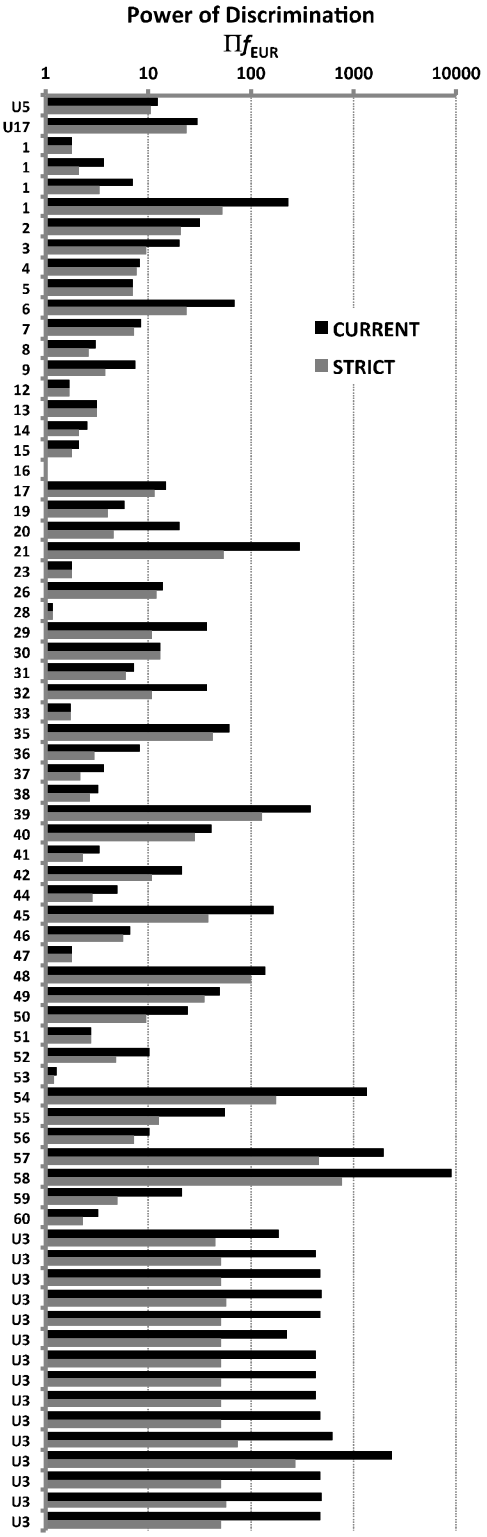
The parameter file in X!Tandem was modified to accommodate universal carboxymethylation of free cysteine residues (C+57), partial oxidation of methionines (M+16), and partial deaminidation of asparagines and glutamines (N+1, Q+1). Two partial cleavages of trypsin are also accommodated.

From the .xml output, all result records with an expect score <= 1.0 are captured. All peptides ranked as "1" (first choice) are then filtered. The list of peptides, NM numbers and genes are then captured and uniquely sorted for each sample. Peptides from '.v1' variant sequence are separated from peptides not from '.v1' sequences. See database method for description of '.v1' sequences.

For all samples, the unique sorted list of all peptides from variant sequences as well as reference sequences is collected. Using this unique set of peptides, search the database to determine if each peptide only occurs in a single gene, isoforms are considered the same gene. A set of unique peptides are then formed for the variant peptides as well as the non-variant peptides.

A list of variant positions for each peptide is pulled from the files used in forming the PBIT database. Rs numbers and frequencies are then matched from the Snp137Common.txt file.

Appendix 9. Overall Profile Frequency Recalculated with Strictest Product Rule Application



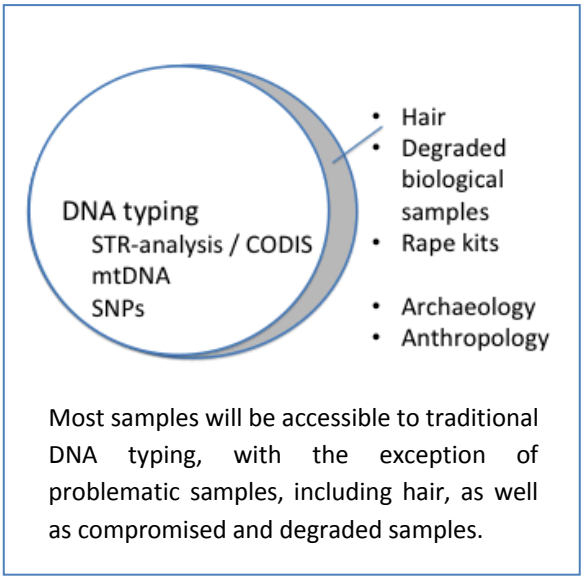
The product rule assumes that the frequency of one locus is not dependent on the status of other loci in the overall genotype. The application of the rule in this study assumed that there was full independence of genotype outside of each gene boundary and full linkage disequilibrium within it (black bars). However, the basic and acidic keratins occur as a cluster on Chromosome 17 and 12 respectively. The analysis of linkage equilibrium within these clusters indicates that there are some examples of linkage disequilibrium between loci on different genes within these clusters, indicating that some modifications to the product rule need to take place. The strictest (and overly conservative) interpretation of the product rule is indicated here (grey bars), where only one locus from within each cluster is used to calculate the overall profile frequency. The maximum overall profile frequency decreases from 1 in 9000 to 1 in 785. On average, an individual profile frequency decreased by a factor of 4 going from the “current” to the “strict” calculation.

Appendix 10. Primer / Factors for Consideration

Background

This project explores the concept of using protein to develop measures of identity and biogeographic background. Current technology focuses on an individual's DNA to calculate these measures. There are good reasons for this: DNA typing is genetically powerful (> 1 in 10^{20}), robust and applicable in many contexts. Due to advances in PCR it is increasingly sensitive. Large legacy databases are based on typing from DNA markers. However, DNA-based methods depend on the presence of intact DNA and DNA is chemically fragile. In the case of compromised samples, hair, or rape kits there is not enough nuclear DNA to provide a full, or even a partial, dataset.

If this is the case the current approach is to try to obtain DNA from the mitochondria. This based on the fact that there are 50 to 1000 copies of mitochondrial DNA in a cell, increasing the probability that there will be enough to obtain identifying information. The problem is that the information is less powerful, giving values of 1 in 5 to 10,000 instead of the > 1 in 10^{20}. The different genetic mechanisms of mitochondrial transmission (maternal line) mean that the conclusions drawn are also different and the information gained may not be actionable. Mitochondrial DNA also suffers from the same chemical processes that adversely affect nuclear DNA levels.



Most samples will be accessible to traditional DNA typing, with the exception of problematic samples, including hair, as well as compromised and degraded samples.

This project proposes another approach: using evidence of genetic variation present in protein. Non-synonymous single nucleotide polymorphisms (nsSNPs) result in a single substitution of an amino acid in the primary protein sequence. This class of genetic variation manifest as a change in the primary structure of a protein, which has the advantage of being more robust than DNA, persisting in the environment after DNA is lost or incomplete. Common nsSNPs also have known allelic and peptide frequencies within a given population, allowing probabilities to be calculated as to whether a given protein sample belongs to a given individual. It may also provide information as to the genetic background of the protein sample.

How are Non-Synonymous SNPs different to other forms of genetic variation?

DNA typing depends on Short-Tandem Repeats (STRs), two copies of which are inherited at any one locus. Each STR has many possible repeat numbers so there are many different options, or alleles, at each locus and any particular outcome will have a low allele frequency. The combination of these, one from each parent, reduces the frequency further. Non-synonymous single nucleotide polymorphisms (nsSNPs) are almost all bi-allelic, with a major allele (allelic frequency > 0.5) and a minor allele (frequency < 0.5). Minor alleles that occur frequently enough to be commonly seen are going to have a

lesser impact on resulting measures of genetic discrimination. Rare alleles that provide higher levels of discrimination necessarily occur at a lower frequency in the population. In either case higher numbers of characterized nsSNP-containing peptides are required to get powerful measures of discrimination. The other main difference is that nsSNPs result in a change in amino acid sequence. This means that the measures of discrimination you achieve are a function of the quality of the proteomic analysis.

How Do Different Typing Methods Compare?

Below is a table comparing and contrasting the two established DNA methods and the protein typing method we examine in this project. The power of discrimination is not as high as DNA, but has the potential to still be very useful in a forensics and intelligence context. Importantly protein, and hair in particular, is very robust and also genetically distinct.

	nsSNP peptides	Nuclear DNA	Mito DNA
Type of Variation	nsSNPs	STR-loci	mtDNA haplotype
Substrate	protein	Nuclear DNA	Mito DNA
Detection	LC/MSMS	PCR/CE	PCR/CE
Genetic Source	Nucleus	Nucleus	Mitochondria
Discrimination Power	1 in 2 to 1 in 30,000	>1 in 10^{20}	1 in 5 to 1 in 10^{4}
Sensitivity in Hair	10 mg	Not fully detected	20 mm of hair

Different approaches to calculate genetic methods of identity.

How does the use of Mass Spectrometry affect our calculations?

Because we detect protein through mass spectrometry of complex peptide mixtures, there is not a guarantee of detection if a peptide bearing an nsSNP is expressed. The only conclusion that can be drawn if an nsSNP-containing peptide is detected is that the individual is not homozygote for the other allele. This results in higher frequencies, a peptide with 20% allelic frequency will occur in 36% of the population (remembering that we have two copies of each gene).

DNA typing depends on detection of PCR products of differing length. Capillary electrophoresis is sensitive and characterized to the point that a single detected allele is enough to conclude that the individual is homozygote. This is not the case with tandem liquid chromatography / mass spectrometry. There are generally around 100,000 different peptides in a complex mixture. In this study we are detecting between 200 and 1,000 peptides. Therefore, absence of a peptide in a resulting dataset cannot be taken as evidence of absence of the corresponding nsSNP in the subject's genome.

Why do we need custom databases?

The other issue resulting from use of mass spectrometry is the need to match spectra with a theoretical sequence. We get two types of information in mass spectrometry, the primary mass of the intact peptide and the spectra of masses resulting from physical fragmentation of the peptide. Most fragmentations occur at peptide bonds and the spectra, therefore, are a characteristic of the primary sequence of the peptide. While a lot of effort has gone into determining a peptide from just the fragmentation spectra, the most reliable method is to compare detected masses with a list of masses from a theoretical peptide database. This means that unknown peptide sequences are absent from the analysis. Therefore, we need to obtain and construct reference databases that contain all nsSNPs. There are some databases that contain variants that are publically available (we use one of them, the GPM database in this study), but these have issues with not being fully complete and have a high level of miss-assignments.